Aviation Research Lab
Institute of Aviation

University of Illinois
at Urbana-Champaign
1 Airport Road
Savoy, Illinois 61874

Supporting Battlefield Situation Assessment Through Attention Guidance and Diagnostic Aiding: A Cost-Benefit and Depth of Processing Analysis

William J. Horrey and
Christopher D. Wickens

Technical Report
ARL-01-16/FED-LAB-01-1

December 2001

Prepared for

U.S. Army Research Laboratory
Interactive Displays Federated Laboratory

Contract DAAL 01-96-2-0003

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **DEC 2001**	2. REPORT TYPE	3. DATES COVERED **-**

4. TITLE AND SUBTITLE **Supporting Battlefield Situation Assessment Through Attention Guidance and Diagnostic Aiding: A Cost-Benefit and Depth of Processing Analysis**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Army Research Laboratory,Aberdeen Proving Ground,MD,21005**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT
Automated support systems may be useful tools for aiding situation assessment in complex environments such as the military battlefield, medical diagnosis, and crisis management. These environments are marked by large amounts of information which often must be weighted and integrated into a meaningful judgment or assessment. Two experiments examined the effects of attention cueing and decision aiding on information integration tasks in static battlefield situations. In the first experiment, sixteen participants completed a resource allocation task for 56 battlefield scenarios (based on perceived threats). For half the trials, an automated system guided their attention to high-relevance information. On 2 trials a memory probe was administered to assess the depth of processing of information, and on the final trial an automation failure was presented. Results demonstrated an overall allocation performance advantage for automation but poorer recall for automation-enhanced units. Half of the participants failed to attend to the system failure. Those participants who detected the failure were inferred to have processed all of the cues more deeply based on their performance on the memory trials. In the second part, 12 participants completed the same task using an automated diagnostic aid (instead of the attention cueing). Again, performance was improved when using automation, more so than in experiment 1. However, there were costs associated with the processing of highly relevant information in these conditions. The costs and benefits of automated cueing and diagnostic aiding are discussed.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**68**	

ABSTRACT

Automated support systems may be useful tools for aiding situation assessment in complex environments such as the military battlefield, medical diagnosis, and crisis management. These environments are marked by large amounts of information which often must be weighted and integrated into a meaningful judgment or assessment. Two experiments examined the effects of attention cueing and decision aiding on information integration tasks in static battlefield situations. In the first experiment, sixteen participants completed a resource allocation task for 56 battlefield scenarios (based on perceived threats). For half the trials, an automated system guided their attention to high-relevance information. On 2 trials a memory probe was administered to assess the depth of processing of information, and on the final trial an automation failure was presented. Results demonstrated an overall allocation performance advantage for automation but poorer recall for automation-enhanced units. Half of the participants failed to attend to the system failure. Those participants who detected the failure were inferred to have processed all of the cues more deeply based on their performance on the memory trials. In the second part, 12 participants completed the same task using an automated diagnostic aid (instead of the attention cueing). Again, performance was improved when using automation, more so than in experiment 1. However, there were costs associated with the processing of highly relevant information in these conditions. The costs and benefits of automated cueing and diagnostic aiding are discussed.

INTRODUCTION

Complex environments are often characterized by large amounts of information as well as multiple dynamic and changing components. A medical doctor must use information from a number of different sources (e.g., various tests, case histories) when making a diagnosis. Air traffic controllers must coordinate many bits of information in order to maintain a complete picture of the current situation. Regardless of the context, each of these components or sources of information may have significant impact on operator decision-making and performance. The extent to which operators in these environments can successfully integrate these sources of information into a coherent situation assessment will directly impact their overall situation awareness as well as their subsequent decisions, actions, and overall performance (e.g., Graham & Matthews, 1999). Though we focus our discussion on the tactical battlefield environment in this paper, the concepts of situation awareness and assessment, information integration, and automation can be readily applied to other domains.

In the battlefield environment, effective commanders must utilize information regarding wide-ranging tactical parameters (e.g., the location of one's own unit in relation to other units (both friendly and enemy); the strength, disposition, and weaknesses of opposing forces; the condition of various avenues of approach), organizational variables (e.g., the level of command; military doctrine; operational orders), environmental factors (e.g., terrain; weather), and various other METT-T (Mission, Enemy, Terrain, Troops, and Time) planning factors (Burba, 1999; Endsley et al., 2000; Evans, 1999). A particularly important component is the reliability of the different sources of information being used in the tactical diagnosis (Wickens et al., 1999; Shattuck et al., 2001). Operations manuals stress that the identification of these and other variables (hazards) is the first step in risk assessment (and the subsequent tactical decisions) (USMC, 1998). Endsley et al. (2000) identify these factors as strong contributors to the establishment and maintenance of situation awareness in the infantry operational environment. As such, commanders' complete and accurate understanding of these factors will impact their perceived tactical risk, subsequent force deployment and protection, and other command and control decisions.

Situation awareness has been the focus of numerous research programs in recent years (see Endsley & Garland, 2000). Endsley's (1995) 3-level model has perhaps been the most frequently cited model of situation awareness (SA). SA involves "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" (p. 36). Level 1 SA involves the perception of cues and elements pertaining to the current situation. These cues are often referred to as the 'raw data' that are available in our surrounding environment (e.g., military reports regarding enemy location and strength; map displays depicting terrain information). Level 2 SA involves the integration and interpretation of the perceived information (from level 1) into a coherent understanding of the current situation (comprehension). The final level (3) of SA involves the projection of current events into the near future (e.g., estimating enemy intent). This level requires a high degree of understanding of the current situational parameters (level 2) and is tightly coupled with operator experience. According to Endsley (2000), SA is considered the main precursor to decision-making, however good SA does not necessarily translate to good decision-making as the latter involves the appropriate weighing of risks and values.

The importance of good SA in the complex, high information battlefield environment (sometimes termed battlefield visualization) is readily apparent and has been acknowledged in the literature and addressed though various research approaches, including display frame of reference and automated decision aiding (see, e.g., Barnes et al., 2001; Thomas et al., 1999; Wickens & Rose, 2001). These currents (SA, information integration, and automated SA aids) provide the framework for the present research.

Information Integration from Multiple Sources

In establishing situation awareness or in any given decision-making or judgment task, people use multiple sources of information to form a hypothesis (or belief in a given hypothesis) regarding the situation or task at hand (Wickens et al., 1999). In many instances the information is derived from qualitatively different sources of information (e.g., radio reports; previous knowledge; map displays). Shattuck and his colleagues (Shattuck et al., 2000, 2001) note that information integration in the battlefield environment will be based largely on contextual factors but also on operational orders, doctrine, and expertise.

As Figure 1 shows, the raw data (or information cues) being used in diagnosis will each have an objective value (or contribution) to the given belief or hypothesis. That is, each cue will have an information value which will bear a specified relationship to the hypothesis, which is a function of the diagnosticity of the cue (the relative importance) and the reliability of the cue (see, e.g., Barnett & Wickens, 1988). The reliability of the cue will depend on a number of factors (e.g., real-world uncertainties, failures in sensors, failures in automation; Wickens et al., 1999). It follows that each cue will vary in its objective information value (see Figure 1), with some cues offering more weight to a given assessment than others (e.g., in the military context, the presence of a nearby enemy is a stronger indicator of a potential attack than is a weather forecast). When observers utilize these cues (integrates them) in making a judgment or assessment, they will impose subjective weights to each cue (based on knowledge or previous experience), which may or may not reflect the true objective values. The means by which an observer uses these cues in making a judgment will vary across individuals and circumstances.

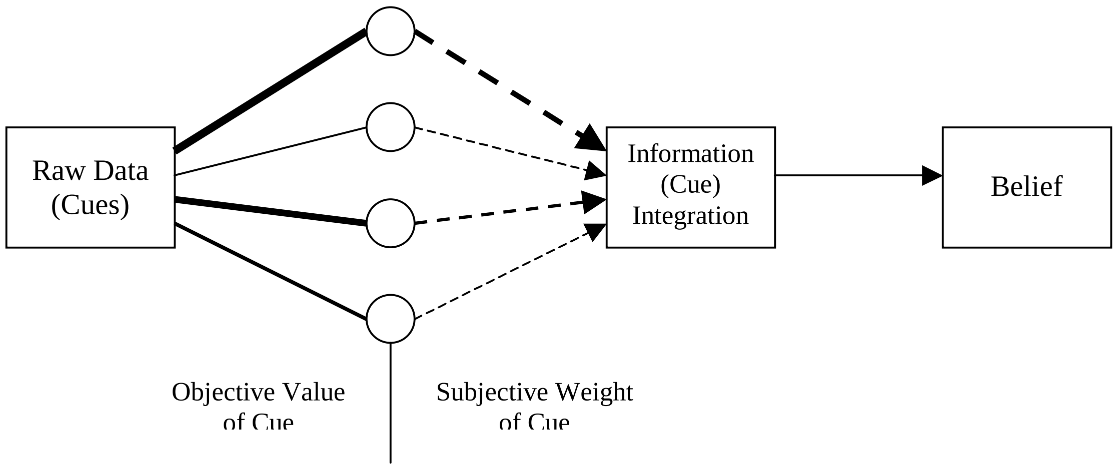


Figure 1. Model of cue integration and belief formation. After Brunswikian lens model (Kirlik, 1995).

This analysis is consistent with the Brunswikian lens model, where a given set of cues bear specified relations to an environmental criterion (to be judged; e.g., judging the threat of an enemy attack; Brunswik, 1952; Hammond, 1966; Kirlik, 1995). The cues (e.g., strength of the enemy force; condition of avenue of approach) and their relation to the criterion will vary as will the ways in which an observer will utilize the cues in making a judgment. Using threat assessment as an example, the cues will contribute differentially to the assessed belief in the outcome "an attack will occur". An observer may utilize these cues in a different fashion, with different weights to arrive at the same (or possibly different) conclusion. The extent to which observers can calibrate (i.e., match their subjective weightings of the cues to their objective values) will determine the overall quality of their SA, judgment, or decision (Wickens et al., 1999). Unfortunately, human observers have limited cognitive, perceptual, and attentional abilities that impact their ability to process large amounts of information. The integration task places high demands on selective and divided attention (attentional resources; Wickens & Carswell, 1995) as well as working memory. In some cases, observers will cope with high cognitive demands by utilizing the pattern of cues to estimate the state of the world. These patterns of diagnosis are linked to expertise and have been labeled recognition-primed decision-making (RPD; Klein, 1989). As such, performance in this cue-integration context will depend, not only on the appropriate calibration of the various information cues to the predicted outcome but also on an observer's ability to allocate attention to various cues accordingly.

People may use heuristics when confronted with difficult cue integration tasks, particularly under time pressure. One example is the 'as if' heuristic by which people will treat differentially weighted cues as if they were of equal value in order to simplify the process of diagnosing a given set of information cues (e.g., Kahneman & Tversky, 1973; Slovic et al., 1977). In many complex environments, this heuristic may have important repercussions. For example, in assessing the likelihood of an enemy attack, a commander may afford enemy strength and accessibility the same relative importance in their tactical assessment where it is inappropriate to do so. Such cognitive short cuts or simplifications may be utilized under conditions of high cognitive load (workload) or time pressure, when there are fewer available resources with which to integrate the relevant pieces of information (Wickens & Hollands, 2000). Research has shown that as the number of information sources increases, people will not typically utilize more than a small subset of cues, even though the extra information could lead to more accurate diagnoses (e.g., Wright, 1974; Dawes & Corrigan, 1974; Dawes, 1979; Schroeder & Benbassat, 1975). Not all research findings have revealed these cognitive shortcuts however. Brehmer and Slovic (1980) examined whether high demand integration tasks would lead to such simplifications in cue-judgment relationships. That is, whether subjective ratings of different cues is distorted in integration tasks. Results from this three-part study did not reveal any evidence for cognitive simplifications. It is possible that the task was sufficiently easy (they used only 2 or 3 cues in their diagnosis, and therefore did not impose a sufficiently high workload) that it did not require any mental shortcuts. Similarly, they did not introduce any time pressure or other potential resource draining tasks (e.g., using distractor items or a secondary task). Under these conditions, we might expect to see degraded performance on information integration tasks and subsequent judgments (Wright, 1974; Svenson & Maule, 1993).

It is generally understood that people will weigh cues differentially and may employ heuristics or mental shortcuts when making a diagnosis or decision. What is less clear is how different cue types impact these two processes. One important issue is the extent to which more abstract (probabilistic) information can be processed compared to more concrete information (e.g., size). Tversky and Kahneman (1981) note that people are often biased in their estimation of probabilistic information. As such, the reliability of the information can be a significant variable in people's ability to integrate cues appropriately. According to models of cue integration, differentially weighted information degrades performance on information integration tasks (consistent with the 'as if' heuristic), including varying degrees of reliability (Sorkin et al., 1991) though not all studies have showed evidence of this degradation (e.g., Jones et al., 1990).

There have been a number of investigations into the effects of unreliable (or uncertain) information on integration. It has been shown that, in some cases, people will suppress the uncertainty of the information as a mechanism to cope with it (Lipshitz & Strauss, 1996, 1997). In an examination of information seeking behavior of U.S. Army enlisted men, Levine and Samet (1973) found that less information is sought when the information is more unreliable in nature. As a result, decision accuracy is greater under conditions of highly reliable information. St. John et al. (2000) had Marines make tactical decisions on information with three levels of reliability. Decision-making in this military context required participants to synthesize information from many different sources (e.g., maps, briefings). The uncertainty of this information was dependent on the source of the information, the reliability of the source, and the age of the information. The results revealed that less experienced Marines elected to "wait and see" (i.e., wait for further information regarding enemy units) under conditions of high uncertainty more often than more experienced soldiers (cf. Levine & Samet, 1973). When information was of medium or low uncertainty, the frequency of “wait and see” decisions was comparable across both experience levels. Using a similar display of information certainty, Kobus et al. (2000) measured decision response times in dynamic tactical scenarios under conditions of low and high uncertainty. Results showed that selection of a course of action (time to acquire SA and make decision) was significantly slower when displayed information was of high uncertainty.

In summary, previous research has shown that limitations in attentional resources and working memory and conditions of high mental workload and unreliable information may lead to degraded performance on information integration and decision-making tasks and, as a consequence, decreased situation awareness. These elements are all a significant part of complex environments, where degraded performance may have serious, life-threatening consequences. By supporting the acquisition and integration of information cues (particularly indices of reliability) or through diagnostic support, technological solutions and various forms of automation may yield positive benefits in this domain and help reduce the cognitive demands of operators and consequently, enhance performance. We now discuss the manner in which automation devices have been designed to provide such assistance, describing their strengths as well as their potential weaknesses.

Automated Systems and their Impact on Performance

Automation involves the execution by a computer (or machine) of a task that was formerly executed by human operators (Parasuraman & Riley, 1997). As such, the definition of automation encompasses a wide range of systems, and stretches also across many domains. For example, future army endeavors will likely incorporate automated systems such as the Army Battle Command System (ABCS) and the Maneuver Control System (MCS) touted at maximizing commander situation awareness through good visualization and integration of information (Burba, 1999).

Automation Taxonomy. Parasuraman et al. (2000) propose a 4-stage taxonomy of human-automation interaction. In this model, automation can be applied (in varying degrees or levels) at any of the stages: (a) information acquisition (attention guidance), (b) information analysis and integration (diagnosis), (c) selection of decision and action (choice), and (d) action implementation. These four stages are based on a simple model of human information processing (sensory processing; perception / working memory; decision making; response selection).

The level of automation applied to each stage of the model will dictate how much control the human is afforded in the operation of the system. Automation in the information acquisition stage (stage 1) acts to support human sensory and attentional processes (e.g., detection of input data). A higher level of automation at this stage may present (on a display) only information it deems appropriate while filtering out all the rest. A lower level of automation, on the other hand, may present all of the raw data but guide attention to what the automation infers to be the most relevant features (target cueing; information highlighting). At the next stage, information analysis, automation serves to aid the human operator by reducing the cognitive demands through use of computer algorithms that may be used to integrate relevant information, draw inferences, and predict future trends. In this stage, lower levels of automation may extrapolate current information and predict future status (e.g., cockpit predictor displays). Higher levels of automation at this stage may reduce information from a number of sources into a single hypothesis regarding the state of the world. At stage 3 automation (selection of decision and action), lower levels may provide users with a complete set (or subset) of alternatives while higher levels may only present the "optimal" decision or action. Finally, stage 4 automation (action implementation) will aid the user in the execution of the selected action.

The model proposed by Parasuraman et al. (2000) maps onto Endsley's model of SA, with early stages of automation contributing to the establishment and maintenance of SA (as shown in Figure 2). It follows that automation in the first stage (information acquisition) that supports the underlying psychological processes of sensation, perception, and attention will also support SA at this early level. Similarly, the extent to which the second stage automation (information analysis) can support cognitive functioning and working memory will directly impact the higher levels of SA.

For all the benefits of automation, there are also limitations and concerns of operator over-reliance upon imperfect automation (Parasuraman & Riley, 1997; Mosier et al., 1998; Moray, 2000; Dzindolet et al. 1999). Endlsey (1996) notes that automation may impact situation awareness through changes in vigilance and monitoring tasks (complacency); changes in operator role from active to passive ('generation effect'; Slameca & Graf, 1978); and changes in the nature of feedback given to the operator. Consistent with these changes in operator roles, Metzger and Parasuraman (in press) demonstrated the detrimental effects of passive versus active monitoring in a simulated air traffic control task.

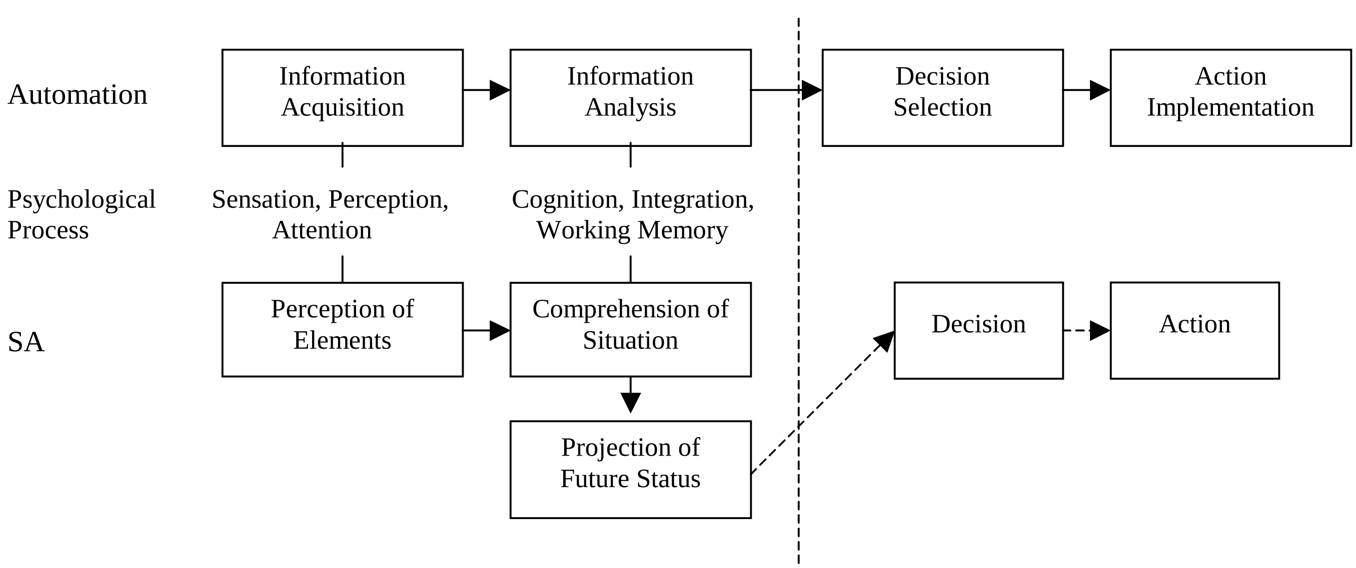


Figure 2. Models of Human Interaction with Automation and Situation Awareness (Parasuraman et al., 2000; Endsley, 1995).

Research has shown that inaccurate decision aids at stages 2 and 3 of automation will affect performance differentially, typically with automation failures at later stages having more serious performance repercussions (e.g., Crocoll & Coury, 1990; Sarter & Schroeder, 2001). Mosier (1997) highlights some key issues in the use of automated decision aids, including the capacity of the user to ignore automation cues in favor of the raw data when appropriate to do so, and the ability to detect failures and errors in automated systems. These issues are of critical concern, especially when the failure of a system has high costs.

Given the potential negative effects of higher-stage automation failure and the importance of strong performance, decision-making, and SA, lower stage (i.e., stage 1) automation may lend itself best to complex environments. A problem with many current systems results from too much source information, creating difficulties finding relevant information at the appropriate times (Endsley, 2000). Through attention guidance, target cueing, and information filtering, early stage automation may help decrease cognitive load but still afford the human observer sufficient autonomy to establish and retain good SA. For example, Evans (1999) emphasizes the importance of automated filtering aids in future battlefield operations. These aids would reduce the amount of information that commanders must consider. Only relevant cues and reports would pass through the filters allowing commanders the capacity to make more effective decisions, especially when under time duress. Recall, this filtering is considered to be higher-level stage 1 automation since the system is selectively hiding some pieces of information. Lower level automation at this stage differs in that the raw data for less relevant sources is available to the user, should they need to consult it.

Benefits of Attention Guidance Automation. There has been extensive research into the effects of stage 1 automation (attention guidance) in target detection tasks. Basic research has reliably demonstrated the capacity for visual cues to reduce search times in target search tasks (see, e.g., Egeth & Yantis, 1997; Flanagan et al., 1998). Applied research has also demonstrated these benefits in military situations (Yeh et al., 1999, Yeh & Wickens, 2001), helicopter hazard detection (Davison & Wickens, 2001), and a number of other domains (Mosier et al., 1998).

Metzger and Parasuraman (2001) examined the benefits of a stage 1 automated aid on conflict detection for air traffic controllers (ATC). The automated aid highlighted a potential loss of separation (conflict) 6 minutes in advance. The aid increased the number of detected conflicts and reduced the search times compared to a non-automated control condition and reduced the controller workload. NASA TLX ratings of workload suggested a slight trend, with higher workload for the manual (non-automated) condition.

In their examination of the effects of perceptual support activities on dynamic decision-making performance, Kirlik et al. (1996) showed that response selection and execution in a simulated football game was faster when participants were provided with visual enhancements of discrete, critical cues. In a subsequent experiment, perceptual support was increased through the enhancement of additional information (including critical properties and relationships). Participants in a battlefield task used an augmented display to assess a number of cues relevant to their combat decisions. The augmented display was used to support the perceptual assessment of these various cues. Performance using this augmented display was superior to the non-augmented display under conditions when workload was increased (by increasing the number of elements in the display).

Stage 1 Automation Costs. Despite these observed benefits from these reported studies, there have been findings which demonstrate potentially negative impacts of reliable automation on the overall processing of information in a display, in particular, the processing of information which is not explicitly highlighted through the automation. In a series of studies that examined the influence of attentional cueing on battlefield target detection, Yeh and her colleagues (Yeh et al., 1999; Merlo et al., 1999; Yeh & Wickens, 2001) found that such cueing narrowed the focus of attention around the cued target such that it reduced the accuracy of detecting more important (uncued) targets that were present in the same scene. Similarly, Davison and Wickens (2001) found that automated cueing of targets (hazards) for helicopter pilots degraded performance in detecting a second, uncued target visible at the same time.

These findings of attentional narrowing have important repercussions for the use of automation, particularly when considering the level of automation being incorporated at early stages. Higher levels of automation in stage 1 will likely filter out the uncued targets (i.e., those that the systems deems unimportant or non-task related) and so, under conditions of perfectly reliable automation, detection performance for these filtered targets will not be relevant. However, in cases where a lower level of automation is adopted (i.e., certain targets are highlighted but not others, as in the studies described above), the extent to which the cued information interferes with the perceptual processing of uncued information may have significant consequences, especially when uncued information has some bearing on the performed task.

In addition to the impact of reliable automation on performance, there are also obvious concerns over the impact of unreliable (or less than perfect) automation. Several studies discussed above, in the context of attentional narrowing, have also addressed the issue of unreliability in early stage automation. Yeh and Wickens (2001) examined the effects of reliable and unreliable target cueing on attention and trust. In this target detection task, targets were cued at either 100% or 75% reliability levels. The target cue consisted of a lock-on reticle that was superimposed over the target item (or when unreliable, the reticle was superimposed over similar looking distractor items). The results demonstrated that cueing information results in a decrease in sensitivity to the features of the raw data (in signal detection sense) suggesting that users were exhibiting an over-reliance on the automation guidance cue, a shift in response criterion, rather than using the cue to increase processing of the raw data underlying the cue. As a consequence, when the cue (unreliably) highlighted a non-target, participants were likely to misclassify it as a target.

In their study of ATC conflict detection aiding, Metzger and Parasuraman (2001) included simulated failure trials in which an aircraft would deviate from its flight plan, thereby creating a conflict. In these conditions, the flight path change was short-term and thus was not reflected in the automated display which cues the controller to likely conflicts. In the automated condition, there were slower response times and a higher miss rate for the failure trials compared to the manual condition. This is consistent with previous findings and the notion of automation induced complacency.

In their examination of helicopter hazard cueing, Davison and Wickens (2001) found that the first occurrence of unreliable cueing resulted in delayed maneuver responses. For subsequent (post-failure) trials, hazard maneuvers were executed earlier than in 100% reliable and baseline conditions, suggesting that pilots' trust in the automated system was reduced after the occurrence of a failure. The impact of failures on user's calibration will likely dictate how frequently the user will employ the system. Miscalibration or undertrust in a system may decrease its overall use, even in situations where it is perfectly reliable.

Stage 2 Automation Costs. Mosier et al. (1998) investigated automation over-reliance in the cockpit of automated aircraft. Over-reliance reflects a miscalibration of user's perceived reliability of the systems and may be characterized by errors resulting from the use of automated cues in lieu of vigilant information seeking and processing of all of the raw data. In this study, pilots flew different flight legs using typical flight deck automated systems. Over the course of these legs, five automation failures were introduced (generated in different automated systems, e.g., flight control system; communications system). Responses to the failure events showed strong evidence of automation over-reliance with pilots failing to utilize all of the available information in making their judgment, attending instead to the highly salient automated cues. Contrary to expectations, pilot experience did not reduce the occurrence of automation over-reliance, rather those pilots with more experience were actually more susceptible to such errors. Automation over-reliance may be a function of systems that are typically highly reliable (e.g., flight deck automated systems). Such biases can have important implications at all stages of Parasuraman et al.'s (2000) taxonomy, especially in high-risk environments such as the cockpit or the battlefield.

Davis and Pritchett (1999) employed a computer-based automated diagnostic tool to aid professional helicopter pilots in diagnosing mid-flight system failures. Throughout 13 flight failures, it provided accurate information, which pilots found beneficial. On the final (14th) failure, the system provided an incorrect diagnosis (and corresponding action recommendation), contraindicated by the raw data. Only 5 of 12 pilots ignored the automation failure and responded appropriately, and 5 others followed the automated guidance, leading to an inappropriate shut down of the remaining good engine.

Wickens et al. (2000) found further evidence for over-reliance on automation inference systems. In this study, pilots flew different flight legs while interacting with a predictive display of traffic (cockpit display of traffic information, CDTI). Pilots tended to over-rely on the automation, allocating more attention to the predictor display than the raw data, especially with increased task complexity.

In general, research on target cueing has demonstrated faster detection times for cued (or highlighted) targets however degraded performance in detecting uncued targets. This degraded performance has significant repercussions for unreliable (or imperfect) automation. Research has shown that observers have slower responses to uncued events when automation is unreliable (in the case of a failure). There is also some indication that the use of target cueing will decrease an observer's sensitivity to or depth of processing of a target (i.e., attending more to the cue than to the raw data underlying the cue). Much of the reviewed research involves target detection and perception tasks (i.e., Endsley's level 1 SA). However little or no research has been done to examine how the implementation of an automated attention guidance system will impact performance on the multi-cue integration task (i.e., Endsley's level 2 SA), characteristic of the commander's formation of battlefield SA.

Automation and Depth of Processing. Target cueing is assumed to modulate the allocation of attention to events or stimuli in the environment. On the one hand, less attention is allocated to uncued targets (attentional narrowing). On the other hand, possibly, less attention is allocated to the raw data underlying the cue (with greater reliance upon the cue itself; Yeh & Wickens, 2001). While this attention modulation can be directly reflected in detection performance, as in the target detection studies described above, its measurement is more challenging in the information integration task examined here, since each event or object does not describe a single "task" whose performance can be assessed. To address this issue, we assume that the depth of processing of each object, modulated by attention, is correspondingly reflected in the memory for the attributes of an object (Craik & Lockhart, 1972). In our paradigm, memory probes may be used to differentiate between the two possible strategies of cue use contrasted by Yeh and Wickens (2001); decreased response bias (increased cue reliance) and increased sensitivity (increased processing of raw data). If observers are adopting a response bias strategy, they would likely exhibit poorer recall for different attributes of the raw data underlying the cue. On the other hand, those who adopt a sensitivity strategy would demonstrate better recall on the same memory task. In our experiment, the conceptual framework proposed by Craik and Lockhart (1972) offers a useful approach for examining depth of processing in an information integration task where multiple pieces of information must be attended to, and are sometimes cued.

Summary and Present Research

As is the case in many other domains, battlefield commanders' situation awareness often involves the integration of large amounts of information from a number of sources in order to form an accurate situation assessment (Graham & Matthews, 1999). This weighted information includes the location and strength of other friendly and opposing forces, the surrounding terrain, and a large number of other METT-T operational factors. Previous work has shown that people do not always integrate multiple pieces of information optimally (when making a judgment or decision), especially under conditions of high workload, time pressure, or when the information is unreliable in nature, conditions which are characteristic of the battlefield environment. Automation can be provided to assist the battlefield commander in this task at various stages of information processing, for example in guiding attention to the most valuable cues (stage 1), in diagnosing what automation infers to be the most likely state of intent (stage 2), or in recommending the most appropriate course of action (stage 3) (Parasuraman et al., 2000). However limitations of automatic diagnosis and choice have been found in operator over-reliance upon imperfect automation (Parasuraman & Riley, 1997; Mosier et al., 1998). Thus, in this experiment we focus our primary interest on automation at the first stage, to assist the operator by highlighting the most relevant cues for situation assessment. Unlike automated situation assessment and choice, this technique does not need to hide the raw data, but only de-emphasizes that which is less relevant. Research on target cueing (a form of attention guidance) has reliably demonstrated the benefits of automation. Nevertheless such highlighting or attention cueing has also been found to produce unwanted effects on attentional tunneling (e.g., Yeh et al., 1999; Metzger & Parsuraman, 2001; Davison & Wickens, 2001), and over-reliance.

While past research on automation attention guidance has focused on target detection tasks (e.g., Yeh et al., 1999; Davison & Wickens, 2001), the current research examines stage 1 attention cueing in an information integration task (i.e., Endlsey's stage 2 SA) where all the raw data are available and the cues highlight the most relevant information (i.e., most highly weighted in integration). Specifically, we assessed the effects of an automated cueing aid in a static battlefield map display on (a) the assessed threat of enemy attack from the east and west, (b) the depth of processing of raw data (for high and low relevant information, cued and uncued), and (c) over-reliance on imperfect automation (the participant's reaction to the automation's failure to cue a highly relevant piece of information).

In two experiments, participants under time pressure observed map displays which contained large amounts of information (regarding the type, location, strength, and accessibility of other military units, as well as the reliability of the information source). In experiment 1 (stage 1 automation), the cueing aid highlighted the enemy units that were most relevant to the participant's threat assessment and was intended to help the observers filter out the less relevant information (e.g., neutral or other friendly units). We hypothesized that the filtering effects of the automated aid would allow participants to make more optimal defensive allocations compared to baseline conditions. Memory probes were used on some trials to assess differential effects of automated cueing on the depth of information processing (Craik & Lockhart, 1972) for a particular unit (i.e., whether cueing decreased target sensitivity) (Yeh & Wickens, 2001). It was also predicted that the failure of automation to highlight a relevant cue would result in the failure to process that cue and hence an inappropriate allocation of resources. Finally, we were interested in whether certain information cue types would be intrinsically given more weight in the threat assessment, independent of the level of automation and their information value (i.e., concrete versus abstract probabilistic cues).

Experiment 2 allowed the differences between stage 1 and stage 2 automation to be examined. In this experiment, an automated diagnostic decision aid (stage 2) replaced the cueing aid that was incorporated into the first part of this research. This decision aid made suggestions regarding the appropriate deployment of defensive resources rather than highlighting relevant information. It was predicted that, to the extent that this higher stage automation was reliable, performance would be superior to the stage 1 automation in the first study. It was also anticipated, however, that the costs associated with the failure of this automation would be greater (as demonstrated by the failure trial) to the extent that participants become over-reliant on the automated aid. Such a finding was postulated by Parasuraman et al. (2000) and would be predicted on the basis of findings by Crocoll and Coury (1990) and Sarter and Schroeder (2001). These studies demonstrated that automation failures at later stages caused greater decrements to performance than those at earlier stages.

EXPERIMENT 1

METHODS

Participants

Ten upper level ROTC students (ages 20-23, M = 21; ROTC experience, M = 3 yrs) and six non-ROTC (graduate) students (ages 23-38, M = 28) at the University of Illinois volunteered for this study. Eleven men and 5 women made up these groups. All participants had normal or corrected-to-normal vision and were familiar with topographical (contour) maps. All participants were paid $7 US per hour for completing the study.

Materials

Hardware. Battlefield scenarios were presented to participants on a 21-inch Silicon Graphics color monitor through a 180 MHz Silicon Graphics O2 workstation with 128 MB of RAM. The monitor was set to 1280 x 1024 pixels of resolution. Battlefield scenarios were created using in-house graphics and development software.

Battlefield Scenarios. Sixty-four battlefield scenarios were developed using topographical maps of Fort Irwin and standard military symbology (USMC, 1997). Four sections of the Fort Irwin region were selected for their varied terrain features. Standard symbols for enemy, neutral, and friendly units were embedded within these map sections (see Appendix A). These units varied in size (e.g., platoon), type (e.g., enemy combat mechanized), location, and the reliability of the intelligence estimate of their identity. Three levels of reliability were used which represented varying degrees of certainty: highly reliable information (confirmed; marked by solid lines), medium reliability (marked by dashed lines), and low reliability (unconfirmed; marked by dotted lines) (see Appendix B). For non-ROTC students, a numerical digit replaced the standard symbology for unit strength. The participant's own unit was always located near the center of the map. Summary information for each scenario is presented in Appendix C.

On each trial, participants had 20 defensive resources which they could deploy to either the east or west of their position. Participants were required to evaluate the overall threat of units in the east versus those in the west and allocate defense resources accordingly. Optimally, a large threat from the east, for example, would receive a larger proportion of these resources than would a lower perceived threat from the west. The overall threat was the sum threat of each individual unit occupying a particular region (all units were operating independently). The relative threat of each unit to the participant's current location was based on weighted evidence from multiple cues. Participants needed to integrate information on unit type and size, the separation distance (relative to their own position), the difficulty of the terrain between the unit and themselves (straight line approaches were specified), and the reliability of the cue.

Automation. On some of the trials, an automation feature was incorporated into the battlefield display. This automation guided attention to the most relevant (highest threat) symbols on the map by augmenting them. Symbols subject to this enhancement pulsed from high to low intensity at a rate of approximately 1 Hz. The relevance of a symbol was based on its information value (units having higher information value were deemed to be more of a threat; Barnett & Wickens, 1988) and this information value was based on several variables (size, type, distance, and difficulty of terrain, and reliability). The following formula, depicting the information value of a particular unit, was derived through a multiple regression of questionnaire data from six independent observers (see Appendix D):

(1) $$\mathbf{IV_{unit} = X_{type}(90 + 4\,X_{size} - 5\,X_{dist} - 14\,X_{diff}) \times R},$$

where, X_{size}, X_{dist}, and X_{diff} define the unit size, distance, and difficulty of the terrain, respectively. R is the overall reliability of the information (from 0 to 1), and X_{type} is the type (1 for enemy units, 0 for neutral or friendly). Four independent observers rated the difficulty of terrain on a 4-point Likert scale (4 being the most difficult terrain). It follows from this formula that only enemy units will be perceived as a threat, and threat increases as unit size increases, separation distance decreases, and terrain difficulty eases. Reliability is used as a moderator variable (see Appendix E, for sample IV calculations). The automation feature enhanced symbols that had information values equal to or greater than 30, yielding on the average trial, automation highlighting of approximately 22% of the units.

Memory Probe. A memory probe was administered following two (roughly 4%) of the scenarios. The purpose of this probe was to determine to extent to which participants were attending to the raw data (the unit symbols). The probe was administered unpredictably and in lieu of the participant's allocation response and queried details on the size of the unit at a particular location in the battlefield display (see Appendix F). Participants gave their confidence rating on a five-point Likert scale. One probe followed a non-automated trial (no enhancement), while another followed an automated trial (queried either an enhanced symbol or a non-enhanced symbol). Responses were scored on the basis of accuracy and degree of confidence.

Failure. One scenario was presented in which the automation feature failed to enhance all of the highly relevant units. On this trial, the enhancement appeared normal for all of the units in one direction however did not highlight a very important unit on the opposite side (one which would have a substantial impact on the allocation of resources). The purpose of this trial was to determine whether participants were attending to all of the raw data on automated trials or rather to the enhanced units only. This element was never the target of a memory probe.

Procedure

Participants completed an informed consent form (Appendix G) and a brief demographic questionnaire (Appendix H) at the beginning of the 45-minute session. Participants were seated at a SGI workstation and given brief verbal instructions (see Appendix I for verbal protocol). This instruction set familiarized the participants with the maps and contour lines, military symbology, rules of engagement, automation features, and task demands.

Participants were instructed to assume the role of a battlefield commander positioned in a central unit. As the commander, they were asked to make critical decisions for the defense of their position based on information obtained from a map display. Participants were instructed to observe each battlefield scenario carefully and rate the relative threat from forces in the east versus those in the west (based on size, type, distance, difficulty of terrain, and reliability). Using this judgment, they were required to allocate 20 defensive resources to the appropriate east-west positions (e.g., 13 east and 7 west). Participants were told that the purpose of the automation was to guide their attention to the most relevant units on the battlefield and that non-highlighted units were not necessarily irrelevant but rather deemed to be *less* of a threat than the highlighted units.

Each trial began with a brief instruction screen after which the battlefield scenario appeared (on keystroke). The trial ended when the participant pressed another key or after 25 seconds had elapsed. This time value was chosen (after pilot testing) to impose considerable time-stress to perform the task accurately, and thereby to assure that the assistance from the automated highlighting was both required and used. The map display then disappeared and the response screen appeared. Participants first completed a brief practice block (5 scenarios) followed by the experimental block, which consisted of 51 scenarios. On roughly half of the trials, the automation feature was active. Automation scenarios were randomly selected and counterbalanced across participants (in a set of four different presentation orders). A memory probe question was administered on two of the trials. On the final trial of the block, participants were presented with the failure trial. The self-paced block was approximately 30 minutes long.

Following the experimental block, participants completed a post-experimental questionnaire (Appendix J) and were remunerated for their participated.

Experimental Design

This experiment utilized a mixed design, with the between variable of Student (ROTC, non-ROTC) and the within variable of Display Type (automation, no automation). All participants were exposed to both display types, however they did not experience all 64 scenarios. Each participant was shown one of four subsets of 51 scenarios (see Appendix K). These subsets were used to reduce the session duration. Memory probe trials were counterbalanced to control for order effects.

RESULTS

Equation (1) was used to compute the optimal allocation of defensive resources based on the sum of the information values for the various units displayed on the map (comparing east versus west). Participant allocation responses were compared to the predicted values and expressed as absolute difference (error) scores in the analyses. As such, smaller difference scores were an indication of more optimal performance.

A total of 5 observations were removed as outliers from the subsequent analyses (i.e., they exceeded 3 standard deviations from the mean). Data from the remaining 749 trials were used in the overall analyses.

Allocation Performance. A two-way ANOVA for Student (ROTC; non-ROTC) and Display Type (automation, no-automation) revealed significant main effects for both variables (Student, $\underline{F}(1, 366) = 4.8$, $\underline{p} = .03$; Display, $\underline{F}(1, 366) = 6.1$, $\underline{p} = .01$). Overall, allocation policies were closer to the optimal level for trials with automation ($\underline{M} = 2.7$) versus those with no automation ($\underline{M} = 3.1$) (see Figure 3). This finding is consistent with the hypothesis that automation would benefit performance on the information integration task.

Non-ROTC (graduate) students were found to have lower error scores ($\underline{M} = 2.6$) than ROTC students ($\underline{M} = 3.0$) (see Figure 3). The Student x Display interaction was not significant ($\underline{F}(1, 366) = .11$, $\underline{p} = .74$), suggesting that both groups benefited equally from automation.

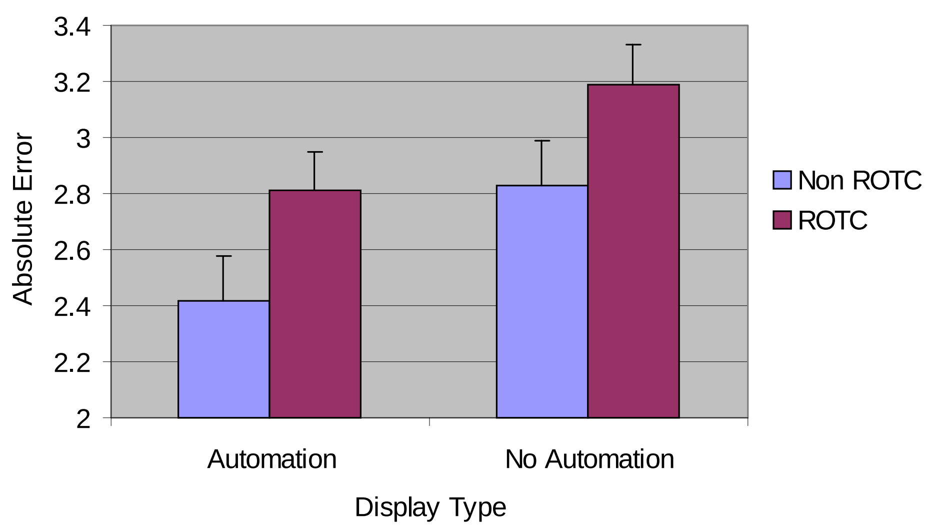


Figure 3. Absolute error by display type and student type

A two-way ANOVA of response times revealed significant main effects for Student ($\underline{F}(1, 374) = 26.8$, $\underline{p} < .001$) and Display Type ($\underline{F}(1, 374) = 16.2$, $\underline{p} < .001$) (see Figure 5). The Student x Display Type interaction was marginally significant ($\underline{F}(1, 374) = 2.7$, $\underline{p} = .10$).

As Figure 4 demonstrates, responses were made more rapidly (M = 18.6 s) on automated trials than on non-automated trials (M = 20.2 s). ROTC students were also found to respond faster (M = 18.5) than non-ROTC students (M = 21.0), thus in conjunction with the accuracy data, suggests that the two groups differed slightly in their speed-accuracy tradeoff.

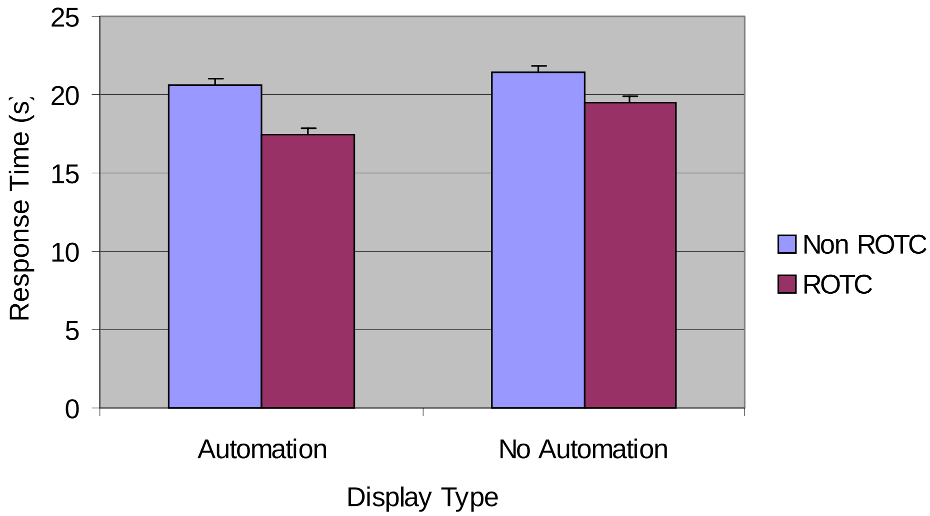


Figure 4. Response time by display type and student type

Memory Probe. A two-way ANOVA was used to determine the depth of processing for high and low relevance units (Relevance) in automation and no-automation conditions (Display Type). Because ROTC and non-ROTC students showed equal benefits of automation and because of the relatively small number of memory probes, this analysis was collapsed across the two groups. Overall, results did not show a main effect for Display type on the confidence-based measure of unit memory ($F(1, 27) = .27$, $p = .61$) nor a Display x Relevance interaction ($F(1, 27) = .27$, $p = .61$). The main effect for Relevance approached significance ($F(1, 27) = 3.5$, $p = .07$) suggesting that participants adopted the appropriate strategy of processing highly important cues more deeply (M = 5.9) than less important ones (M = 4.2).

In conditions with no automation, recall performance for the high relevance unit (M = 6.5) was higher than for the lower relevance unit (M = 4.2) (see Table 1). This difference was marginally significant, ($F(1, 14) = 3.4$, $p = .09$), and suggests that observers, under normal (non-automated) conditions, are appropriately attending to objects that are more important to their threat assessment task. For automated conditions, this trend favoring recall for the high relevance unit, which was highlighted (M = 5.5), over the low relevance unit, which was not (M = 4.2), was still present however much weaker (non-significant; $F(1, 13) = .75$, $p = .40$). The introduction of automation appears to have some adverse effect on the depth of processing for high relevant, enhanced units. This finding offers support for the notion of degraded sensitivity or less processing of the raw data within highlighted or cued targets (Yeh & Wickens, 2001).

Display Type	Object Type	
	Low Relevance	High Relevance
Automation	4.2 (1.4)	5.5 (.78)
No Automation	4.2 (.6)	6.5 (1.3)

Table 1. Recall scores for memory probe questions (Std. Error in parentheses).

Memory performance for the low relevance objects was equal, regardless of automation condition. Analyses of the raw scores indicated that performance for these units was above chance performance. Because the unit was not highlighted in both of these conditions, this suggests that the depth of processing for these cues was not hindered by the presence of automation for other items. This finding is not consistent with the findings from other research that the presence of cued targets detracts attention from non-cued objects (e.g., Yeh et al., 1999; Yeh & Wickens, 2001).

As noted above, recall for the high relevance item was slightly weaker with automation ($\underline{M}$ = 5.5) compared to the no automation ($\underline{M}$ = 6.5) condition. Performance for this high-relevant (automation highlighted) memory probe was characterized by a bimodal distribution, with participants typically scoring either very high or very low in the automated condition (see Figure 5). The resulting high variance in this response pattern barred any significant findings, but is of considerable interest in its own right suggesting that some participants may have ignored the raw data behind the highlighted cue entirely, integrating only the fact of its highlighting, whereas others used the highlighting as a guide for deeper analysis of the threat that had been highlighted. These two strategies correspond to the effects of cueing that Yeh and Wickens (2001) had associated with reliance, or response bias (beta) and enhanced processing, or sensitivity (d’), respectively.

Failure trial. On the failure trial, the automation did not cue a highly relevant target. In this scenario, the perception of this unit was designed to have a significant impact on the allocation of defensive resources. Thus, whether a participant noticed the unit or not was inferred from their allocation score for this trial, using an experimenter-defined criterion to make this inference. This criterion was based on the optimal allocation of resources when the uncued target was taken into consideration. Scores that did not fall within 2 points of this criterion level were considered to be an indication that the unit had not been noticed and / or was not utilized in the allocation of resources. Results suggested that roughly half of the participants (7 of 15) failed to notice the high-relevant unit that the automation did not highlight. This relatively high figure may be an indicator of automation-induced complacency. In the post-experimental questionnaire, some “noticers” noted that the automation missed some important enemy units, while some “non-noticers” commented on the automation’s capacity to make them ignore the non-highlighted information.

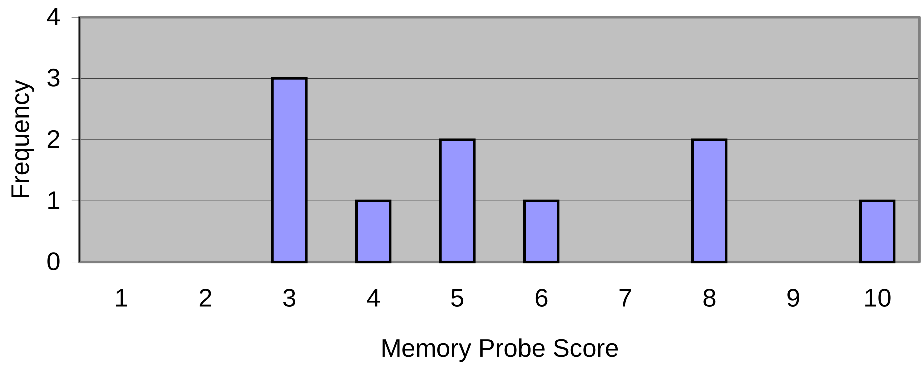


Figure 5. Memore probe response frequency for automated, high-relevant cue

Given the frequency of missed events on the failure trial, we examined whether there was any significant relationship between performance on the failure trial and the pattern of responses on the memory probe for the high-relevance, cued target, as shown in Figure 5. The presence of such a relationship would perhaps offer an explanation for the observed bimodal probe response pattern. A point biserial correlation between observer type (failure noticer, non-noticer) and performance on the memory probe revealed a significant relationship ($\underline{r}_{pb}$ = .69, $\underline{p}$ < .05) between the two variables. It was estimated that 63% of the variance in memory probe performance was accounted for by observer type. There was however no indication that demographic variables (e.g., gender, ROTC vs. non ROTC students) might distinguish between the two observer types.

Further examination of observer type revealed some interesting trends with respect to the unit relevance of the memory probe (high, low), though statistical tests were precluded due to low cell counts. As shown in Table 2, noticers and non-noticers tend to perform equally on recall for low relevance units on non-automated trials. However when automation is present, noticers scored much higher than non-noticers on the low-relevant (and therefore uncued) objects, suggesting that the two groups engaged in different strategies for interacting with the automation. This is consistent with the results from the failure trial, where noticers were more likely to attend and react to a non-highlighted unit.

For the high relevance memory probe, noticers again outperformed non-noticers on recall for unit attributes (see Table 2). The relatively high score for non-noticers on the non-automated, high relevant probe may be an indicator that these observers perform well in general but this performance degrades when automation cueing is introduced.

Memory Probe	Display Type	Observer Type	
		Noticer	Non-noticer
Low Relevance	Automation	7.5 (1.5)	2.0 (.6)
	No Automation	4.2 (1.0)	4.3 (.6)
High Relevance	Automation	6.7 (1.0)	3.8 (.5)
	No Automation	4.0 (-)	9.3 (.3)

Table 2. Recall scores for low and high relevance memory probe (Std. error in parentheses).

Cue Weighting. Several analyses were carried out to investigate the differential treatment of cue types in allocation responses. Specifically, we were interested in determining how observer's judgments were influenced by differences in unit strength, distance, terrain, and reliability information. In order to accomplish this, several scenarios were matched to allow for comparisons across these dimensions. Pairs of trials were compared in which differences along one of the dimensions (e.g., reliability) required a different allocation policy (for these trials, all other cue dimensions were held constant). If participants did not attend to changes in the particular cue, then we would expect their response patterns to be similar on the two trials (i.e., no difference). As such, the difference in the allocation scores between the two trials was used in the following analyses. The expected difference for optimal allocation for the selected trials was between 3.4 and 4 for each of the four different cue types.

Preliminary analyses were run to determine whether observers attended to changes along one of the dimensions. These initial analyses compared the difference scores (for the two trials) against zero (i.e., the expected response if they did not attend to the change). Tests for each of these variable were found to be significant: unit size ($t(45) = 9.6$, $p < .001$); reliability ($t(24) = 4.9$, $p < .001$); terrain ($t(69) = 9.9$, $p < .001$); and distance ($t(93) = 10.7$, $p < .001$). These tests demonstrate that participants were indeed attending, at least to some extent, to each of the four cue categories (as reflected by their response patterns).

A one-way ANOVA for Cue Type on non-automated trials revealed significant differences across Cue Type ($F(2, 80) = 4.3$, $p = .02$). (The distance cue was not included in this analysis because of the potential confound with terrain difficulty. These two properties are inexorably linked and therefore highly correlated within the map display, and though steps were taken to minimize these influences, it was nearly impossible to control for all terrain types while manipulating distance values). Post hoc tests showed a greater influence of size ($M = 5.3$) than terrain ($M = 2.8$; $p = .01$) and reliability ($M = 3.7$; $p = .16$) though the latter difference was only marginally significant (see Figure 6). The difference between reliability and terrain was not significant ($p = .29$).

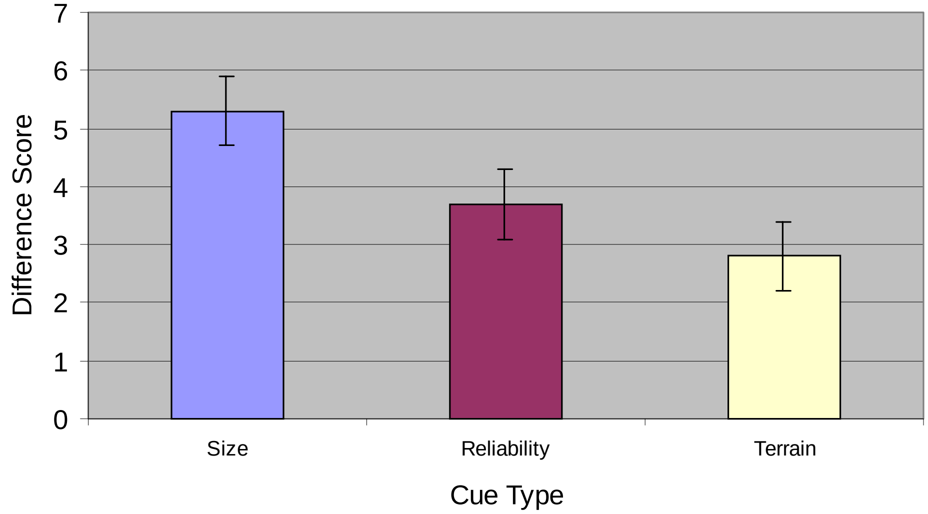


Figure 6. Difference scores by cue type for non-automated trials

The rank order of cue influence (size > reliability > terrain) that was inferred from the objective performance data on non-automated trials is not entirely consistent with subjective self-reported importance, as measured in the post-experimental questionnaire. Participants indicated that size was the most important factor ($\underline{M}$ = 4.4), followed by distance ($\underline{M}$ = 4.0), terrain ($\underline{M}$ = 3.9), and reliability ($\underline{M}$ = 3.2) (see Figure 7). Non-parametric rank tests indicated that this subjective ordering was significant ($\underline{X}^2_F$ = 13.7, $\underline{p}$ = .008) and somewhat consistent across raters (Kendall's coefficient of concordance = .19).

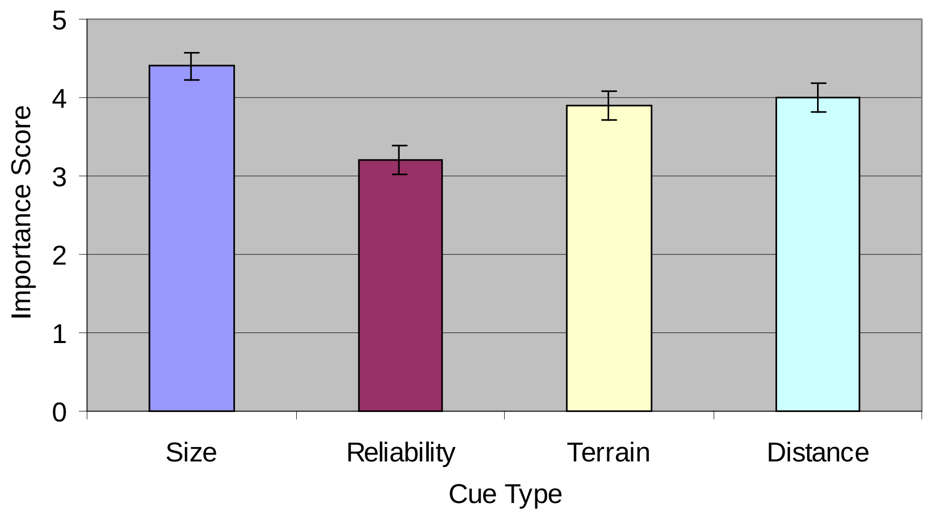


Figure 7. Self-Reported Importance by Cue Type

Questionnaire responses. In general, participants found the automated cueing aid to be moderately useful (M = 3.3 on a 5-point Likert scale) and had many positive comments regarding the potential for such systems (see Appendix L, for participant responses). Many participants lauded the ability of the system to help them quickly detect the most threatening units in the map display, as well as their ability to filter out the less relevant stimuli. They also acknowledged a number of different situations where the system would be most useful, including conditions of time pressure, and high workload (from number of sources of information).

Interestingly, participants were also aware of many potential shortcomings of the system, including: the presentation of unreliable information or automation failure; the capacity of the system to detract attention from uncued hazards (attentional tunneling); and discrepancies between the computer’s assessment of threat and their own.

DISCUSSION

The goal of the present experiment was to examine the impact of stage 1 attention cueing on a battlefield integration task. While most research on early stage automation has focused on the detection of cued targets (as a primary task), this study cued targets of relevance to be integrated in forming a situation assessment and a subsequent allocation decision. That is, the stage 1 automation used in the present study supported a level 2 SA task (Endsley, 1995). While the primary performance measure reflected this level 2 SA (error scores), two converging operations were employed to infer the impact of the automation on individual cue processing; the depth of processing memory probe and the failure catch trial.

Automation costs and benefits

Primary Task Performance. In the non-automated condition, performance on the allocation task was reasonable, suggesting that there was some processing of the numerous information cues in the time available. However overall, performance with the automated cueing aid was superior to unaided performance, with reduced departures from the optimal allocation scores in automated conditions. The response times with the aid were 1.5 seconds shorter than for the non-automated conditions suggesting that automation allowed the participants to make more speeded and accurate allocation decisions, presumably by allocating their attention (visual search) initially to the cued items. Alternatively, by reducing the perceptual demands of visual search, the automation may have availed more cognitive resources for the information component of the task (Liu & Wickens, 1992). In general, this finding is consistent with previous research on reliable target cueing (e.g., Yeh et al., 1999; Davison & Wickens, 2001), however it extends beyond simple detection tasks to higher-level integration tasks.

Depth of Processing. Results from the memory probe revealed a difference in recall for high-relevant versus low-relevant units. The improved recall for the more important objects suggests that observers are processing these cues more deeply than less important ones. This follows intuitively and lends support to the memory probe as an ecologically valid measure of the depth of processing (Craik & Lockhart, 1972). Such better recall and deeper processing also explain the degree of optimality of the resource allocation scores.

Previous research has shown that the presence of cued targets detracts attention from other uncued targets (e.g., Davison & Wickens, 2001; Yeh et al., 1999; Yeh & Wickens, 2001). This finding was not replicated in the present study. Recall scores for the low-relevant (uncued) units were equal in both the automated and non-automated conditions but still above chance performance, suggesting that the presence of automated cues did not have an adverse impact on processing of these units. The inconsistencies in the impact of automation on uncued targets between this and prior research may be due, in part, to the nature of the tasks employed. As mentioned previously, most research has utilized target detection tasks (level 1 SA) to demonstrate the tunneling of attention around cued target locations, thus each cue could be processed independently of other cues. The current study, however, required participants to integrate multiple pieces of information (level 2 SA), a many-to-one mapping of cues to task performance. Furthermore, the amount of reduction in RT achieved by the cueing, 1.5 seconds, was sufficiently small to suggest that it did not eliminate inspection of the uncued items altogether, a conclusion also supported by the above-chance accuracy of memory for those uncued items.

Recall for the attributes of the high-relevant unit exhibited a somewhat different pattern of results. The general (non-significant) trend showed inferior recall in the automated condition compared to the baseline condition, suggesting that the application of automated cueing to these high importance targets may negatively impact the depth of processing for these cues. More important was the evidence of a bimodal response pattern in the recall scores for the cued high relevance units. This pattern suggests that different observers adopted different strategies for interacting with the stage 1 automation. Observers who had poor recall for the cued target may have failed to attend much to the raw data present in the display, attending primarily to the highlighting. For example, they may have noted the presence of 2 cued targets in the west and 4 cued targets in the east and proceeded to allocate twice as many resources to the east without processing these cues at a deeper level. Yeh and Wickens (2001) found a similar response bias (beta) in observers who believed the automated system to be highly reliable. In their target search study, participants were found to attend more to the information suggested by the presence of the cue rather than to the raw data underlying it.

In contrast, the observers who exhibited good recall in the present study may have been using the cueing to direct their attention to the relevant features for deeper analyses. This strategy would suggest an increase in sensitivity (d’) to the information in the cued target, an effect that was also observed by Yeh and Wickens (2001). No differences were found however to suggest a demographic variable which could account for the observer type. Are there any implications of these differing beta and d’ strategies in the use of automation? The former (beta shift) may be a more efficient strategy under time pressure however there will be costs if automation is unreliable, an issue we now address.

Failure Trial. The failure trial exhibited some degree of evidence for automation induced complacency or over-reliance. Roughly half of the participants failed to notice the automation failure (an uncued, but highly relevant item) and hence made inappropriate allocation responses. On all trials prior to the failure trial, the automation had operated reliably, consistently highlighting the most relevant units. Over-reliance and complacency are an unfortunate negative by-product of highly (yet imperfect) automated systems (Parasuraman & Riley, 1997; Mosier et al., 1998; Metzger & Parasuraman, in press). As such, the appropriate level of human interaction with such systems must be clarified to ensure safe and efficient use of automation (Bainbridge, 1983).

A significant finding relating to the failure trial is the relationships between noticing the uncued high-relevant unit in this trial and scores on the memory probe measured on earlier trials. These relationships lend further support to the notion that there are different (beta and d') strategies for interacting with the automation. Some observers will utilize the automation to get a global sense of the situation and make their response on the basis of this high-level assessment. This strategy reduces the cognitive demands of the integration task and, given the performance findings, often leads to good allocation decisions. However it is in cases where detailed information needs to be recalled or when automation is unreliable that this advantage breaks down. Alternatively, observers may attend to the local highlighting cues, inspecting each in turn.

While the d' strategy just described would directly predict an enhanced ability to notice that a cued item was *not* of high relevance (i.e., an automation cueing "false alarm"), it is important to realize that the automation failures employed here (and better detected by the "noticers") was of the opposite type: an automation cueing "miss". Thus the quality of deeper cue processing showed by the noticers must have applied to both cued and uncued items alike, as their performance on the low relevance memory probe would suggest. Subsequent analysis revealed that this differential strategy neither slowed nor speeded the overall RT, compared to the non-noticers.

It would be beneficial to have some measure of eye movements in order to better explore these different strategies. Such measures would reveal any differences in visual search patterns when observers view the map displays. In their examination of ATC conflict detection, Metzger and Parasuraman (2001) found that observers who did not notice the automation failure event had fewer fixations and shorter dwell times than those who detected it, suggesting the presence of different visual scan strategies for interaction with automated systems.

The presence of such different strategies may have important implications in real-world design and applications. The nature and conditions of the task will likely dictate which strategy is more appropriate. For instance, under time pressure adopting a beta strategy (i.e., trust the cues) may be appropriate given that overall allocation performance in the automated conditions was good. When time pressure is not significant, when a task demands recall for specific target details, or when automation is unreliable or imperfect then a d’ strategy may be the best strategy. In order for automated systems to accrue their intended benefits, users must understand how to interact with the system appropriately, an end which may be attained through training or feedback implementation.

Cue Weighting. These analyses suggested that observer’s judgments were influenced differentially by differences in unit size, terrain, and reliability of information. Both objective and subjective measures indicated that unit size information had a more significant impact on allocation responses than the latter cues. The military symbol (or numerical digit) for unit size was a highly concrete information cue, which may have contributed to the strong influence on response patterns. The terrain cue, though a concrete (physical, geographical) feature itself, was found to be less influential perhaps because the use of this cue required the observer to integrate information about the enemy unit and contour lines with information regarding the position of one’s own unit (hence, increasing mental workload). Reliability, in contrast, is a more abstract cue than the concrete size and terrain cues. That is, reliability is a probabilistic information cue, which is often subject to biases in estimation (Tversky & Kahneman, 1981)**,** and not always effectively used in judgments (Wickens, Gordon & Liu, 1997). The current findings did not suggest any difference in cue influence between reliability (abstract probabilistic) and terrain (concrete) cues perhaps due to the graphic display of three different levels of reliability. This graphic display may have reduced the abstractness of the cue, allowing observers to treat it as if it were a concrete cue.

While certain benefits and costs of stage 1 automation (Parasuraman et al., 2000) are expressed in this research, it is less clearly understood how higher stages of automation involving automatic diagnosis will impact performance in the battlefield scenario. The second study examined stage 2 automation (diagnosis) in the same experimental paradigm, such that the costs and benefits of these two stages of automation could be directly compared. Parasuraman et al. (2000) suggest that progressively later stages of automation, by reducing the amount of cognitive work, can produce greater performance benefits if the automation is fully reliable. However, a possible implication is that the costs of unreliability might also be amplified at later stages, a finding observed by Sarter and Schroeder (2001) when stages 2 and 3 were compared. The current study appears to be the first one to contrast stages 1 and 2 within the same paradigm.

EXPERIMENT 2

METHODS

Participants

Twelve students at the University of Illinois volunteered for this second study (ages 22-33, M = 26). Six men and 6 women made up this group. All participants had normal or corrected-to-normal vision and were familiar with topographical (contour) maps. All participants were paid $7 US per hour for completing the study.

Materials

The experimental set up and battlefield scenarios in this study were the same as those employed in the first phase of this research.

Stage 2 Automation. In contrast to the attention guidance automation used in Experiment 1, this experiment used stage 2 automation. Rather than cueing the most relevant (highest threat) units, the automation suggested an appropriate allocation response. There was no stage 1 automation (target highlighting) in this part of the study. On a given automated trial, two red boxes containing the suggested allocation appeared to the east and west of the participant's unit (see Appendix M). This suggestion was based on the optimal allocation as determined by Equation (1).

Memory Probe & Failure. The memory probe trials were similar to those administered in Experiment 1. Probes queried size attributes of high- and low-threat units in both automated and non-automated conditions. In this experiment, high-threat units were not enhanced in the automated condition.

The failure scenario differed from that employed in Experiment 1. In this phase, the automation suggested an inappropriate allocation for the displayed units. This suggestion failed to consider a very important unit in one direction. The purpose of this trial was to determine whether participants were attending to all of the raw data on automated trials or rather on the automated aid alone. This element was never the target of a memory probe.

Procedure

This study followed the same procedure as described in Experiment 1. Participants were instructed that the computer's assessment was only a suggestion and that the final allocation decision would be theirs to make. They were told that the automation was highly reliable but not perfect (see Appendix N for the revised verbal protocol and Appendices O and P for the revised questionnaires). Several scenarios were excluded from this phase of the research because some of the displayed units overlapped with the automated aid. The experimental block consisted of 43 trials, including the 2 memory probes and 1 failure trial.

RESULTS

As in the first phase of this research, absolute difference (error) scores (between the predicted and participant's allocation) were used in the analyses, with smaller difference scores indicating more optimal performance.

Allocation Performance. A one-way ANOVA on allocation error revealed a significant effect for Display Type (automation, no automation; $F(1, 221) = 39.8$, $p < .001$). Overall, allocation scores were improved with the automated aid ($M = 1.7$) compared to without ($M = 3.0$). This finding is consistent with the hypothesis that reliable stage 2 automation would benefit performance on an information integration task.

An ANOVA for response time did not reveal any significant differences between the Display conditions ($F(1, 221) = .94$, $p = .33$).

Memory Probe. A two-way ANOVA was used to determine the depth of processing for high and low relevance units (Relevance) in automation and no-automation conditions (Display Type). The results revealed main effects for Unit Relevance ($F(1, 20) = 7.0$, $p = .02$) and Display Type ($F(1, 20) = 8.5$, $p = .009$) (see Figure 8). Scores for the high relevance unit were higher ($M = 5.7$) than for the low relevance unit ($M = 4.1$), suggesting that participants appropriately attended more closely to the highly important cues. Recall performance with the automated aid was degraded ($M = 4.0$) compared to the unaided condition ($M = 5.8$), suggesting that the presence of the automation reduced the likelihood of processing the cues more deeply.

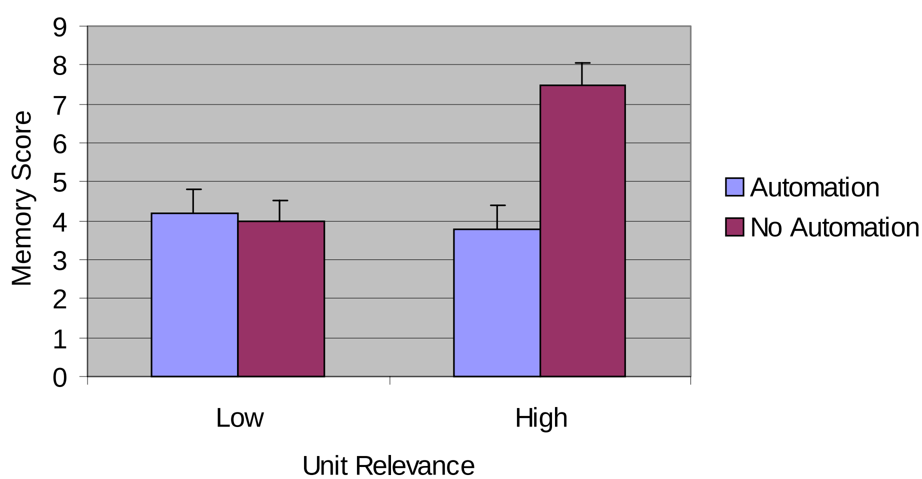


Figure 8. Memory probe scores by unit relevance and display type

Both main effects can best be interpreted within the context of the significant Relevance x Display interaction ($\underline{F}(1, 20) = 10.2$, $\underline{p} = .005$). As shown in Figure 8, memory probe performance in automated conditions was comparable for both the high and low relevance units ($\underline{M}$ = 3.8 and 4.2, respectively). Furthermore, the memory probe scores for the low relevance units was equal, independent of automation level. When no automation was present, recall for the high relevance unit was higher ($\underline{M} = 7.5$) than for the low relevance unit ($\underline{M} = 4.0$) suggesting that the presence of the automation led to less processing of attributes only for the most relevant raw data.

A two-way ANOVA for memory probe response times did not reveal any main effects of Relevance ($\underline{F}(1, 20) = 1.6$, $\underline{p} = .22$) nor Display Type ($\underline{F}(1, 20) = .38$, $\underline{p} = .54$). There was, however, a significant two-way interaction ($\underline{F}(1, 20) = 8.6$, $\underline{p} = .008$) of the same general form as for accuracy, suggesting a mild speed-accuracy tradeoff (see Figure 9). In automated conditions, participants took longer to respond for the low relevance probe ($\underline{M} = 22.2$) than for the high relevance probe ($\underline{M} = 19.7$) whereas in the no automation condition, the pattern was reversed (low, $\underline{M} = 18.7$; high, $\underline{M} = 25.1$). This pattern of response times may offer some explanation for the observed memory probe scores for these conditions, with higher scores being associated with increased response times. That is, quite intuitively, deeper processing requires more time to accomplish.

Failure trial. On the failure trial, the automation made an inappropriate suggestion, one that did not consider the presence of a very important unit. The inclusion of this unit would have significantly altered the suggested values. Whether a participant noticed the unit or not was inferred using the same criterion as in experiment 1. Results suggested that roughly half of the participants (5 of 11) failed to notice the high-relevant unit or noticed it but opted to allocate their resources according to the automation's suggestion. This finding is consistent with the findings in the first experiment.

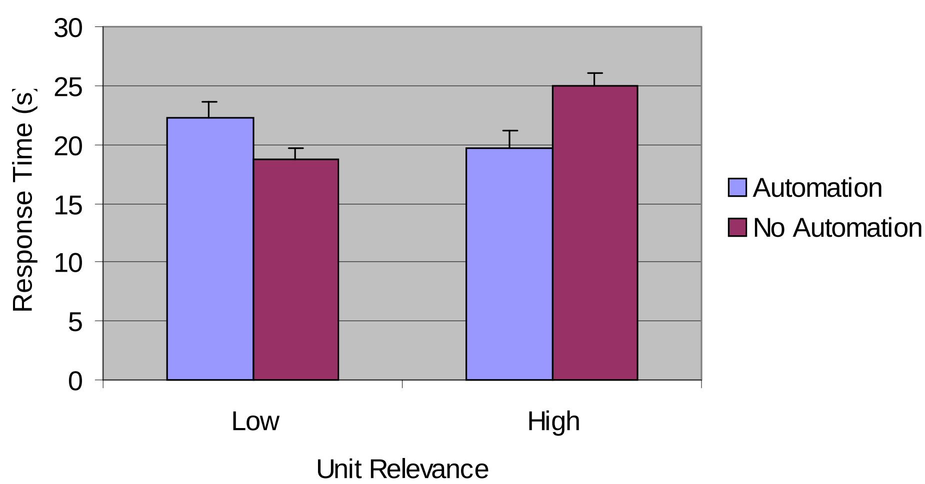


Figure 9. Response times for memory probe by unit relevance and display type

Questionnaire responses. Participants rated the automated aid as being moderately useful (M = 2.8 on a 5-point Likert scale) and had mixed comments and criticisms regarding the application of such systems (see Appendix Q, for participant responses). Time pressure and uncertainty were touted as situations where the automated aid would be beneficial. Many participants appreciated the fact that the aid could act as a second opinion for diagnosing the situation or as a baseline for reaching a decision. Many expressed concerns, however, over the fact that they did not understand how the aid reached it's recommendations or that it sometimes did not agree with their own allocation decision.

Stage 1 versus stage 2 automation. As shown in Figure 10, the different stages of automation employed in Experiments 1 and 2 yielded different benefits in performance, as expressed in percent reduction in error. The application of stage 1 automation (attention cueing) helped reduce allocation error by 13%, while stage 2 automation (diagnosis) contributed to a 43% reduction in error.

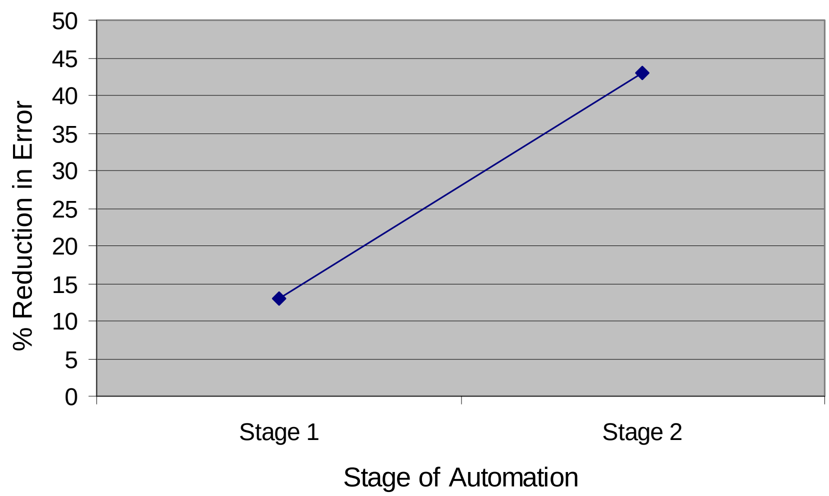


Figure 10. Percent reduction in error by automation stage

Figure 11 compares the performance on the memory probes across the two experiments. Recall performance was similar for low-relevant units for both stages of automation (as well as non automated conditions). For the high-relevant units, however, performance was worse for the higher stage (2) automation compared to the low stage (1), though both automation types were poorer than baseline conditions.

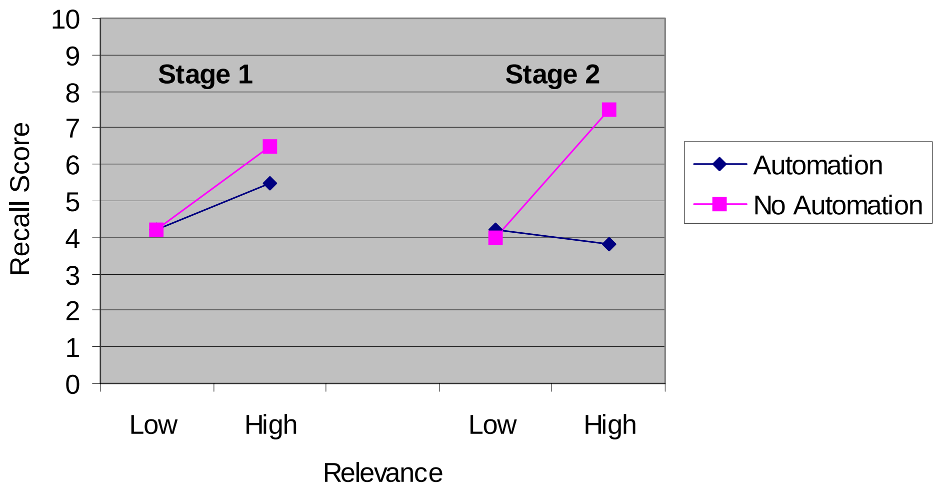


Figure 11. Memory probe score by relevance and stage of automation.

The allocation error scores on the failure trial from the first and second study were compared in order to determine whether there were greater costs associated with unreliable stage 2 automation versus unreliable stage 1 automation. Positive error scores indicated that observers noticed the high-relevant, uncued target whereas negative scores indicated a failure to attend to this unit. A t-test did not reveal a significant difference across the stages of automation ($\underline{t}(24) = .20$, $\underline{p} = .84$), though stage 2 automation had slightly higher costs ($\underline{M} = -.09$) than stage 1 ($\underline{M} = .13$). Thus while the higher stage of automation did, as predicted, lead to shallower processing of highly relevant data than the lower stage, such a difference was not seen, in the current results, to have implications for a poorer response to unreliability.

DISCUSSION

Experiment 2 examined the impact of stage 2 automation (diagnosis) on the battlefield integration task, with the general purpose of comparing the costs and benefits of stage 1 and stage 2 automation.

Automation costs and benefits

The results revealed that allocation performance with reliable stage 2 automation was superior to unaided performance. Optimal performance was moderated to the extent that observers trusted and relied on the automation's suggestions. Alternatively, the automated aid provided observers with a starting point (or "ballpark" figure) for making their own assessment of the situation. The equivalence in response time across display conditions would seem to offer support for the latter. It would be expected that, if participants were to rely solely on the automation's guidance, the response times would be reduced compared to non-automated trials though response accuracy would not be degraded (on reliable trials).

Depth of Processing. As in the first study, participants seemed to be appropriately attending more to units of higher relevance in the non-automated conditions. However, performance was degraded on automated trials, regardless of relevance level (recall for both high and low relevance units was comparable). This is consistent with the hypotheses that the presence of automation will reduce the depth of processing for different cue information. The response times for these trials would seem to suggest a mild speed-accuracy tradeoff, with participants scoring higher on trials when more time was spent observing the map display.

As in experiment 1, it appears that the benefits provided by automation also produced some costs when the automation was imperfect, with roughly half of the participants apparently failing to notice the incorrect automated diagnosis, as inferred by their allocation scores.

GENERAL DISCUSSION

The purpose of Experiment 2 was to allow for some general comparisons of the benefits and costs of stage 1 and 2 automation on an information integration task. Performance on the defense allocation task was superior with both stages of automation compared to the baseline (non-automated) conditions. In study 1, there was a 13% reduction in error when the attention guidance automation was included in the battlefield scenarios. However in the second study, there was a 43% reduction in errors when the automated diagnostic aid provided allocation suggestions. This difference is consistent with the notion that higher stage automation, when reliable, will improve human operator performance. In experiment 1, the cognitive integration needed to be accomplished manually. In experiment 2, this process was carried out by the automation, reducing the cognitive demands placed on the operator.

The downside of highly reliable automation is the potential for users to become over-reliant on it (Parasuraman & Riley, 1997; Wickens, 2000). While the greater benefits of higher stage automation were clearly expressed in the current research, the associated greater costs with higher (than lower) stage automation were not as clear. There was a small performance decrement for unreliable stage 2 automation relative to stage 1, however this difference was non-significant. This cost analysis, however, was based on a single failure trial. It is possible that an examination of automation failures with greater statistical power (including different failure types) would yield stronger support for the automation-performance tradeoff described above.

Recall performance on the memory probes suggests that cue attributes of high-relevant items are processed more deeply with lower stages of automation (stage 1). As noted above, this is consistent with stage 1 automation requiring the operator in this paradigm to accomplish the cognitive integration manually, thereby increasing the likelihood that high-relevant raw data will be attended. This did not appear to extend to low-relevant items. Recall performance suggested that these low relevance cues were processed equally across the two experiments.

Implications

In general, it has been shown that stage 1 and 2 automation have associated costs and benefits for performance on information integration tasks. It is less clearly understood how higher stages of automation involving decision selection and action will impact performance in such information integration tasks, the impact of repeated failures on trust and system use, or the impact of a highly reliable system (long term) on complacency.

The presence of different strategies for interacting with early-stage automation may also have a significant impact on our understanding of human interaction with automation. It is generally accepted that human performance will vary across different levels and stages of automation (Parasuraman et al., 2000). The current research, however, suggests that there can be wide variations in human performance at the same level and stage of automation, depending on how the automation is used by the operator. This makes it more difficult to predict both user performance, as well as the impact of imperfect (or unreliable) automation. As was demonstrated by the failure trial and the memory probe, different interaction strategies may influence the extent to which users will notice automation failures and their ability to recall task-related details (depth of processing of the raw data). Understanding these strategies represents a non-trivial problem because they will likely vary, not only across systems and tasks but also at different stages and levels of automation within the same system. These strategies will have a significant impact on the design and extent of automated systems and, in turn, their task-specific training programs, which may bear a direct influence on the type of strategy a user will employ.

The rapid advance of computer technology dictates that automated systems will be even more widespread in the near-distant future. In the battlefield context, performance with such systems will be a function of integrated observations (visual and contextual) and judgments, as well as automated information (Serfaty, 1999). Such endeavors must strive to assess and incorporate critical elements of battlefield situation awareness (via experts, manuals, and doctrine) and their relative mission-related importance in order to be of measurable success (Serfaty, 1999). Potential threats to SA aids include terrain and weather interference, computer viruses, electronic jamming, spectral interference, electromagnetic pulse systems, and anti-satellite technologies (Evans, 1999). However, despite these technological and environmental concerns, the overall utility of these systems will be linked fundamentally to the human component.

ACKNOWLEDGEMENT

The authors thank Sharon Yeakel for her time and programming skills, and two anonymous reviewers for their comments and feedback. Research funding for this project was provided through a grant from the U.S. Army Research Laboratory under the Federated Laboratory Program, Cooperative Agreement DAAL0196-2-0003. Any opinions, findings, and conclusions or recommendations expressed in this document are those of the authors and do not necessarily reflect the views of the U.S. Army Research Laboratory. The authors wish to thank LTC Keith Beurskens of the University of Illinois Army ROTC detachment, for making participants available.

REFERENCES

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19(6)*, 775-779.

Barnes, M.J., Pierce, L.G., Wickens, C.D., Dzindolet, M.T., & Rozenblit, J.W. (2001). Human Performance Issues in Battlefield Visualization. In M.E. Benedict (Ed.), *Proceedings of the Fourth Annual Federated Laboratory Symposium on Advanced Displays and Interactive Displays*. University Park, MD: US Army Research Laboratory.

Barnett, B.J. & Wickens, C.D. (1988). Display proximity in multicue information integration: The benefit of boxes. *Human Factors, 30(1)*, 15-24.

Brehmer, B. & Slovic, P. (1980). Information integration in multiple-cue judgments. *Journal of Experimental Psychology: Human Perception and Performance, 6(2)*, 302-308.

Brunswik, E. (1952). The conceptual framework of psychology. *International Encyclopedia of Unified Science, 1(10)*, 1-102.

Burba, E.H., Jr. (1999). Army XXI Insights. In S.E. Graham & M.D. Matthews (Eds.), *Infantry Situation Awareness: Papers from the 1998 Infantry Situation Awareness Workshop* (pp. 11-18). Alexandria, VA: US Army Research Institute.

Craik, F.I.M. & Lockhart, R.S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684.

Crocoll, W.M. & Coury, B.G. (1990). Status or recommendation: Selecting the type of information for decision aiding. *Proceedings of the 34th Annual Meeting of the Human Factors Society* (pp. 1524-1528). Santa Monica, CA: Human Factors Society.

Davis, D.D. & Pritchett, A.R. (1999). Alerting system assertiveness, knowledge, and over-reliance. *Journal of Information Technology Impact, 1(3)*, 119-143.

Davison, H.J. & Wickens, C.D. (2001). Rotorcraft hazard cueing: The effects on attention and trust. *Proceedings of the 11th International Symposium on Aviation Psychology*. Columbus, OH: Ohio State University.

Dawes, R.M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571-582.

Dawes, R.M. & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95-106.

Dzindolet, M.T., Pierce, L.G., Beck, H.P., & Dawe, L.A. (1999). Misuse and disuse of automated aids. *Proceedings of the 43rd Annual Meeting of the Human Factors and Ergonomics Society* (pp. 339-343). Santa Monica, CA: Human Factors and Ergonomics Society.

Egeth, H.E. & Yantis, S. (1997). Visual attention: Control, representation, and time course. *Annual Review of Psychology, 48*, 269-297.

Endsley, M.R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37(1)*, 32-64.

Endsley, M.R. (1996). Automation and situation awareness. In R. Parasuraman & M. Mouloua (Eds.), *Automation and Human Performance: Theory and Applications* (pp. 163-181). Mahwah, NJ: Lawrence Erlbaum.

Endsley, M.R. (2000). Theoretical underpinnings of situation awareness: A critical review. In M.R. Endsley & D.J. Garland (Eds.), *Situation Awareness Analysis and Measurement*. Mahwah, NJ: Lawrence Erlbaum.

Endsley, M.R., & Garland, D.J. (2000). *Situation Awareness Analysis and Measurement*. Mahwah, NJ: Lawrence Erlbaum.

Endsley, M.R., Holder, L.D., Leibrecht, B.C., Garland, D.J., Wampler, R.L., & Matthews, M.D. (2000). *Modeling and Measuring Situation Awareness in the Infantry Operational Environment* (Report DASW01-99-D-0013). Alexandria, VA: US Army Research Institute.

Evans, K.L. (1999). Group 2 summary: Situation awareness requirements for platoons, companies, and battalions. In S.E. Graham & M.D. Matthews (Eds.), *Infantry Situation Awareness: Papers from the 1998 Infantry Situation Awareness Workshop* (pp. 71-86). Alexandria, VA: US Army Research Institute.

Flanagan, P., McAnally, K.I., Martin, R.L., Meehan, J.W., & Oldfield, S.R. (1998). Aurally and visually guided visual search in a virtual environment. *Human Factors, 40(3)*, 461-468.

Graham, S.E. & Matthews, M.D. (1999). *Infantry Situation Awareness: Papers from the 1998 Infantry Situation Awareness Workshop*. Alexandria, VA: US Army Research Institute.

Hammond, K.R. (1966). *The Psychology of Egon Brunswik.* New York: Holt, Rinehart, & Winston.

Jones, P.M., Wickens, C.D., & Deutsch, S.J. (1990). The display of multivariate information: An experimental study of an information integration task. *Human Performance, 3(1)*, 1-17.

Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 251-273.

Kirlik, A. (1995). Requirements for psychological models to support design: Toward ecological task analysis. In J. Flach, P. Hancock, J. Caird, & K. Vincente (Eds.), *Global Perspectives on the Ecology of Human-Machine Systems* (pp. 68-120). Hillsdale, NJ: Lawrence Erlbaum.

Kirlik, A., Walker, N., Fisk, A.D., & Nagel, K. (1996). Supporting perception in the service of dynamic decision making. *Human Factors, 38(2)*, 288-299.

Klein, G. (1989). Recognition primed decision making. *Advances in Man-Machine Systems Research, 5*, 47-92.

Kobus, D.A., Proctor, S., Bank, T.E., & Holste, S.T. (2000). *Decision-making in a dynamic environment: The effects of experience and information uncertainty* (Technical Report 1832). San Diego, CA: SPAWAR Systems Center.

Levine, J.M. & Samet, M.G. (1973). Information seeking with multiple sources of conflicting and unreliable information. *Human Factors, 15(4)*, 407-419.

Lipshitz, R. & Strauss, O. (1996). How decision-makers cope with uncertainty. *Proceedings of the 40th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 189-193). Santa Monica, CA: Human Factors and Ergonomics Society.

Lipshitz, R. & Strauss, O. (1997). Coping with uncertainty: A naturalistic decision-making analysis. *Organizational Behavior and Human Decision Processes, 69(2)*, 149-163.

Liu, Y. & Wickens, C.D. (1992). Visual scanning with or without spatial uncertainty and divided and selective attention. *Acta Psychologica, 79*, 131-153.

Merlo, J.L., Wickens, C.D., & Yeh, M. (1999). *Effect of Reliability on Cue Effectiveness and Display Signaling* (Technical Report ARL-99-4/ARMY-FED-LAB-99-3). Savoy, IL: University of Illinois, Aviation Research Laboratory.

Metzger, U. & Parasuraman, R. (2001). Conflict detection aids for air traffic controllers in free flight: Effects of reliable and failure modes on performance and eye movements. *Proceedings of the 11th International Symposium on Aviation Psychology*. Columbus, OH: Ohio State University.

Metzger, U. & Parasuraman, R. (in press). The role of air traffic controller in future air traffic management: An empirical study of active control versus passive monitoring. *Human Factors, 43(4)*.

Moray, N. (2000). Are observers ever really complacent when monitoring automated systems? *Proceedings of the IEA 2000 / HFES 2000 Congress* (vol. 1, pp. 592-595). Santa Monica, CA: Human Factors and Ergonomics Society.

Mosier, K.L. (1997). Myths of expert decision making and automated decision aids. In C.E. Zsambok & G. Klein (Eds.) *Naturalistic Decision Making* (pp. 319-330). Mahwah, NJ: Lawrence Erlbaum.

Mosier, K.L., Skitka, L.J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. *International Journal of Aviation Psychology, 8(1),* 47-63.

Parasuraman, R. & Riley, V. (1997). Humans and automation: Use, misuse, disuse, and abuse. *Human Factors, 39(2)*, 230-253.

Parasuraman, R., Sheridan, T.B., & Wickens, C.D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, 30(3),* 286-297.

Sarter, N.B. & Schroeder, B. (2001). Supporting decision-making and action selection under time pressure and uncertainty: The case of inflight icing. *Proceedings of the 11th International Symposium on Aviation Psychology*. Columbus, OH: Ohio State University.

Schroeder, R.G. & Benbassat, D. (1975). An experimental evaluation of the relationship of uncertainty to information used by decision makers. *Decision Sciences, 6*, 556-567.

Serfaty, D. (1999). Model-based measurement of situation awareness in dynamic tactical environments. In S.E. Graham & M.D. Matthews (Eds.), *Infantry Situation Awareness: Papers from the 1998 Infantry Situation Awareness Workshop* (pp. 173-184). Alexandria, VA: US Army Research Institute.

Shattuck, L., Graham, J., Merlo, J., & Hah, S. (2000). Cognitive integration: An investigation of how expert and novice commanders process battlefield data. In M.E. Benedict (Ed.), *Proceedings of the Fourth Annual Federated Laboratory Symposium on Advanced Displays and Interactive Displays* (pp. 47-51). University Park, MD: US Army Research Laboratory.

Shattuck, L., Merlo, J., & Graham, J. (2001). Cognitive integration: Exploring performance differences across varying types of military operations. In P.N. Rose (Ed.), *Proceedings of the Fifth Annual Federated Laboratory Symposium on Advanced Displays and Interactive Displays* (pp. 199-204). College Park, MD: US Army Research Laboratory.

Slameca, N.J. & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 592-604.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology, 28*, 1-39.

Sorkin, R.D., Mabry, T.R., Weldon, M.S., & Elvers, G. (1991). Integration of information from multiple element displays. *Organizational Behavior and Human Decision Processes, 49*, 167-187.

St. John, M., Callan, J., Proctor, S., & Holste, S.T. (2000). *Tactical decision-making under uncertainty: Experiments I and II* (Technical Report 1821). San Diego, CA: SPAWAR Systems Center.

Svenson, S., & Maule, A. (1993). *Time Pressure and Stress in Human Judgment and Decision Making*. New York: Plenum.

Thomas, L.C., Wickens, C.D., & Merlo, J. (1999). *Immersion and Battlefield Visualization: Frame of Reference Effects on Navigation Tasks and Cognitive Tunneling* (Technical Report ARL-99-3/FED-LAB-99-2). Savoy, IL: University of Illinois, Aviation Research Laboratory.

Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 185*, 1124-1131.

USMC (1997). *Operational Terms and Graphics* (Field Manual 101-5-1). Washington, DC: Department of the Army.

USMC (1998). *Risk Management* (Field Manual No. 100-14). Washington, DC: Department of the Army.

Wickens, C.D. (2000). *Imperfect and Unreliable Automation and its Implications for Attention Allocation, Information Access and Situation Awareness* (Technical Report ARL-00-10/NASA-00-2). Savoy, IL: University of Illinois, Aviation Research Laboratory.

Wickens, C.D. & Carswell, C.M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors, 37(3)*, 473-494.

Wickens, C.D., Gempler, K., & Morphew, M.E. (2000). Workload and reliability of predictor displays in aircraft traffic avoidance. *Transportation Human Factors, 2(2)*, 99-126.

Wickens, C.D., Gordon, S.E., & Liu, Y. (1997). *An Introduction to Human Factors Engineering*. New York: Addison Wesley Longman.

Wickens, C.D. & Hollands, J.G. (2000). *Engineering Psychology and Human Performance* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Wickens, C.D., Pringle, H.L., & Merlo, J. (1999). *Integration of Information Sources of Varying Weights: The Effect of Display Features and Attention Cueing*. (Technical Report ARL-99-2/FEDLAB-99-1). Savoy, IL: University of Illinois, Aviation Research Laboratory.

Wickens, C.D. & Rose, P. (2001). Army Federated Laboratory on interactive and advanced displays and controls. *Human Factors Handbook*. Washington, DC: US Government Printing Office.

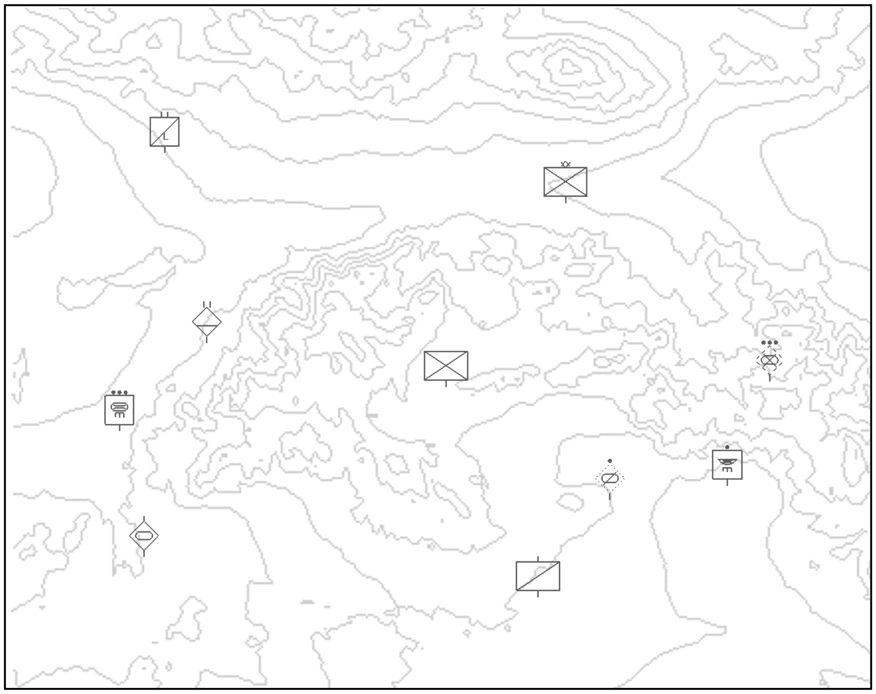
Wright, P. (1974). The harassed decision maker: Time pressures, distractions, and the use of evidence. *Journal of Applied Psychology, 59*, 555-561.

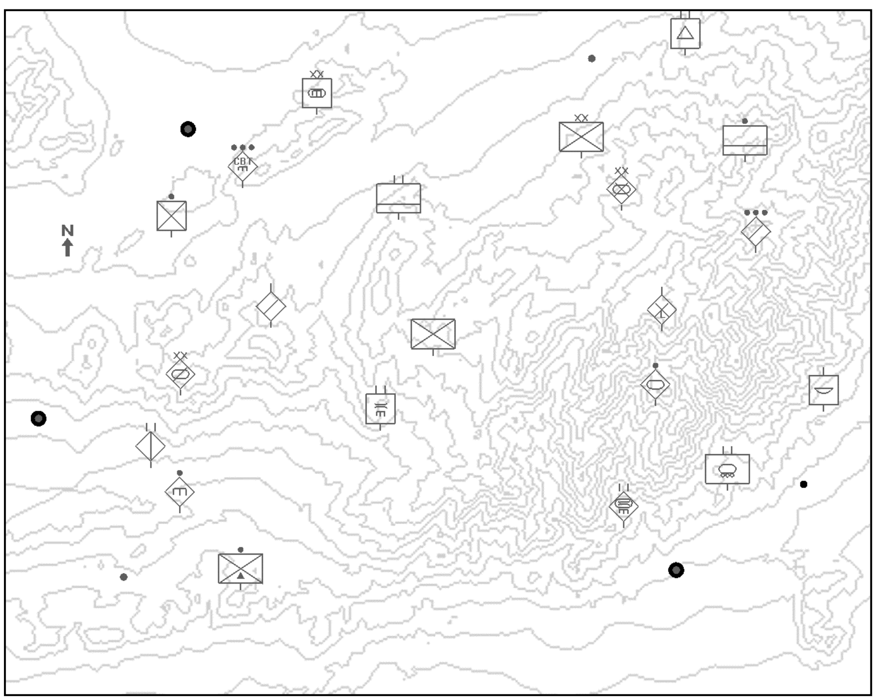
Yeh, M. & Wickens, C.D. (2001) Explicit and implicit display signaling in augmented reality: The effects of cue reliability, image realism, and interactivity on attention allocation and trust calibration. *Human Factors, 43(3)*.

Yeh, M., Wickens, C.D., & Seagull, F.J. (1999). Target cueing in visual search: The effects of conformality and display location on the allocation of visual attention. *Human Factors, 41(4)*, 524-542.

Appendix A

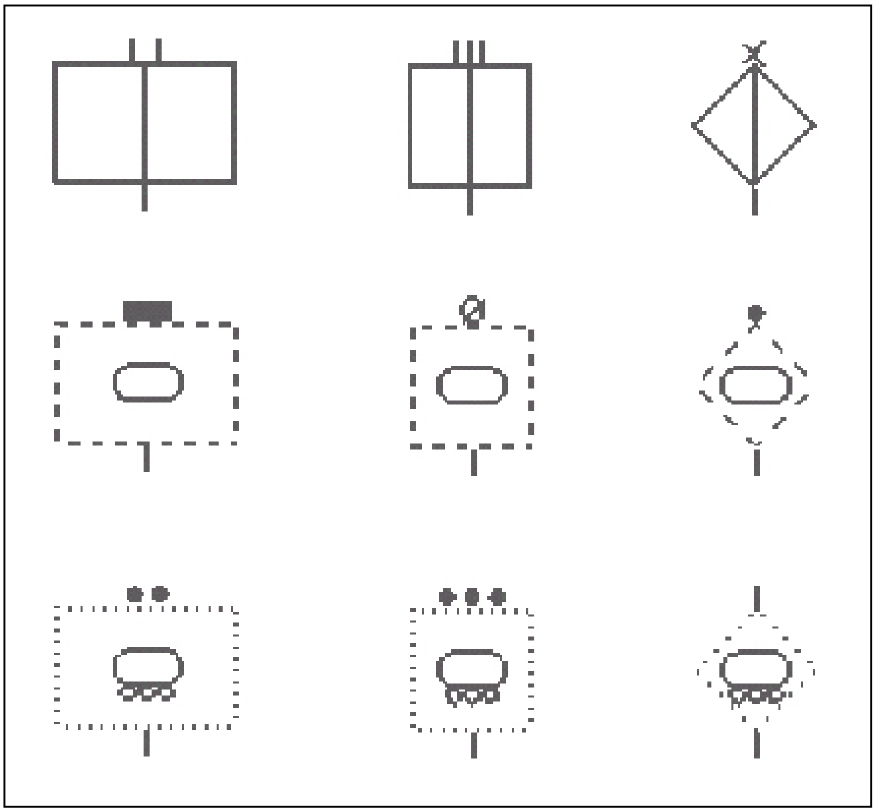
Samples of battlefield scenarios





Appendix B

Reliability levels of military units



High Reliability

Medium Reliability

Low Reliability

Appendix C

Summary information for battlefield scenarios

Scenario	# Enemy Units	$\Sigma\ IV_{West}$	$\Sigma\ IV_{East}$	Optimal allocation (East)	# Units with IV > 30 (W / E)
1	10	24	128	17	0 / 3
2	10	59	208	16	1 / 4
3	10	54	22	6	0 / 0
4	10	62	76	11	1 / 0
5	10	95	177	13	1 / 4
6	10	157	178	11	3 / 3
7	10	151	166	10	2 / 3
8	10	148	148	10	3 / 2
9	10	110	47	6	2 / 0
10	10	180	152	9	3 / 3
11	10	184	169	10	3 / 3
12	10	61	31	7	1 / 0
13	10	101	97	10	1 / 1
14	10	159	184	11	3 / 4
15	10	92	108	11	1 / 1
16	10	184	52	4	3 / 0
17	10	96	197	13	1 / 4
18	10	72	188	14	0 / 3
19	10	139	111	9	2 / 1
20	10	88	33	5	1 / 0
21	10	183	118	8	3 / 1
22	10	232	181	9	4 / 4
23	10	214	133	8	5 / 2
24	10	69	62	9	1 / 1
25	10	172	174	10	3 / 3
26	10	9	79	18	0 / 0
27	10	193	210	10	4 / 4
28	10	93	92	10	2 / 1
29	10	106	128	11	2 / 2
30	10	130	164	11	2 / 3
31	10	120	47	6	2 / 0
32	10	88	131	12	1 / 3
33	10	87	74	9	1 / 1
34	10	84	171	13	1 / 2
35	10	62	122	13	1 / 0
36	10	253	144	7	3 / 3
37	10	126	38	5	3 / 0
38	10	160	217	12	4 / 4
39	10	182	170	10	4 / 3
40	10	60	33	7	0 / 0
41	10	178	97	7	4 / 2
42	10	131	125	10	2 / 2
43	10	31	209	17	0 / 5
44	10	200	171	9	4 / 2
45	10	113	54	6	3 / 0
46	10	116	199	13	1 / 4

47	10	147	170	11	3 / 2
48	10	113	101	9	2 / 2
49	10	80	62	9	2 / 0
50	10	180	209	11	4 / 3
51	10	174	76	6	4 / 1
52	10	52	108	14	1 / 2
53	10	112	62	7	2 / 1
54	10	97	122	11	2 / 2
55	10	183	215	11	3 / 5
56	10	152	158	10	3 / 2
57	10	133	31	4	2 / 0
58	10	275	199	8	4 / 5
59	10	119	194	12	1 / 3
60	10	105	85	9	1 / 1
61	10	200	71	5	4 / 0
62	10	23	117	17	0 / 2
63	10	130	172	11	1 / 3
64	10	251	256	10	5 / 4

Appendix D

Independent observer ratings

Six independent observers were shown a series of maps depicting their own unit and enemy units varying in size, distance, and (terrain) difficulty of approach. Observers were asked to rate the relative threat of each of these enemy units on a ten-point scale (1 = Low Treat, 10 = High Threat). Raters were instructed to base their assessment only on size, distance, and terrain.

Threat scores were collected for 21 different enemy configurations. The median scores for each configuration were used as the criterion variable in a multiple regression. Predictor variables were the size of the unit (as presented on the map), distance (measured in cm from observer's own unit), and the terrain difficulty (as rated on a four-point scale by a different set of four observers):

$$\textbf{Threat Score} = 90 + 4\,X_{size} - 5\,X_{dist} - 14\,X_{diff}$$

This formula was modified to account for the type of unit (i.e., enemy vs. neutral or friendly) and the reliability of the information (R) to yield the following (Threat Score has been renamed Information Value of a particular unit):

(1) $$IV_{unit} = X_{type}(90 + 4\,X_{size} - 5\,X_{dist} - 14\,X_{diff}) \times R,$$

where, X_{size}, X_{dist}, and X_{diff} define the unit size, distance, and difficulty of the terrain, respectively. R is the overall reliability of the information (from 0 to 1, where R<1 denotes degraded levels of reliability), and X_{type} is the type (1 for enemy units, 0 for neutral or friendly).

Appendix E

Sample information value calculations

A) Unit: Enemy Armored Cavalry, Platoon size, 6 cm (map scale) distance, easy terrain, highly reliable information.

	X_{type}	X_{size}	X_{dist}	X_{diff}	R
Unit	1	4	6	1	.85

$$IV_{unit} = X_{type}(90 + 4\ X_{size} - 5\ X_{dist} - 14\ X_{diff}) \times R$$

$$= 1\ (90 + 4(4) - 5(6) - 14(1)) \times (.85)$$

$$= 54$$

B) Unit: Enemy Combat Dismounted, Battalion size, 8 cm (map scale) distance, most difficult terrain, moderately reliable information.

	X_{type}	X_{size}	X_{dist}	X_{diff}	R
Unit	1	6	8	4	.50

$$IV_{unit} = X_{type}(90 + 4\ X_{size} - 5\ X_{dist} - 14\ X_{diff}) \times R$$

$$= 1\ (90 + 4(6) - 5(8) - 14(4)) \times (.50)$$

$$= 10$$

C) Unit: Neutral Light Infantry, Division size, 10 cm (map scale) distance, easy terrain, highly reliable information.

	X_{type}	X_{size}	X_{dist}	X_{diff}	R
Unit	0	10	10	1	.85

$$IV_{unit} = X_{type}(90 + 4\ X_{size} - 5\ X_{dist} - 14\ X_{diff}) \times R$$

$$= 0\ (90+ 4(10) - 5(10) - 14(1)) \times (.85)$$

$$= 0$$

Appendix F

Memory probe questions and confidence scale

1) (A) No automation (B) With automation

   What size was the enemy unit located in the southeast quadrant of the previous battlefield display?

2) (A) No automation (B) With automation

   What size was the enemy unit located in the northeast quadrant of the previous battlefield display?

Unit Size

Squad	Platoon	Company	Battalion	Division
1	2	3	4	5

Confidence Scale

Not at all confident		Somewhat confident		Very confident
1	2	3	4	5

Appendix G

Informed Consent Form

Research Project Title: Supporting Situation Assessment Through Attention Guidance

Investigator(s): William J. Horrey and Dr. Christopher D. Wickens

Description of Research Project:

The purpose of this study is to examine automation in battlefield decision-making. The goal is to gain a better understanding as to how information is integrated and processed. Such knowledge may help in the development of decision aids or assessment tools which will help reduce command and control decision difficulty on the digitized battlefield. For this study, you will be shown electronic maps of battlefields and asked to make some defense decisions. The study should take no more than 60 minutes to complete. If, at any point during the course of this study, you feel uncomfortable you are free to leave without penalty. For completing the study you will receive $7.

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research project and agree to participate. In no way does this waive your legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. You are free to not answer specific items or questions in interviews or on questionnaires. You are free to withdraw from the study at any time without penalty. Your continued participation should be as informed as your initial consent, so you should feel free to ask for clarification or new information throughout your participation. If you have further questions concerning matters related to this research, please contact:

William J. Horrey, Department of Psychology, University of Illinois
Phone: (217) 244-4461, horrey@s.psych.uiuc.edu

Dr. Christopher D. Wickens, Department of Psychology, University of Illinois
Phone: (217) 244-8617, cwickens@s.psych.uiuc.edu

______________________________ _________________________
Participant Date

______________________________ _________________________
Investigator Date

Appendix H

Pre-Experimental Questionnaire: Study 1

Participant _____

1. Age _________
2. Gender _______
3. How much ROTC experience do you have? _______ (months / years)
4. Do you have normal or corrected-to-normal vision? Yes No

Please rate your experience with (or understanding of) the following:

	Little or none		Moderate		Very high
5. Contour maps	1	2	3	4	5
6. Military symbology (e.g., unit size, type)	1	2	3	4	5

Appendix I

Verbal Protocol: Study 1

General Instructions

Provide participant with informed consent.

Thank you very much for participating in this study. It should take approximately 75 minutes to complete. I would like to remind you that you are free to withdraw from this study at any time. Please look through the informed consent.

Participant reads / signs forms.

Do you have any questions?

To begin with, I will ask you to fill out this brief questionnaire. It will ask you a few background questions.

Participant fills out questionnaire.

Today I will show you some electronic maps of battlefields. Your unit is positioned in the center of the map. Your task will be to observe the other units in the area and decide from which direction an enemy attack is more likely. Based on this decision, you will allocate your defensive resources accordingly.

I'd like to over some of the things you'll need to pay attention to when making your assessment. We have attempted to match the symbols to the standard military ones you may be familiar with. Here is a small sample of symbols used here (*Show instruction image 1*).

First of all, note the colour and shape of the symbols. Enemy units are marked by DIAMONDS, neutral units are marked by SQUARES, and friendly units are marked by RECTANGLES. You'll note that inside of each shape is a unit type (e.g., light infantry or engineers – combat dismounted). For the purposes of this study, this unit type will not be important and can be ignored, only whether or not the unit is an enemy, neutral, or friendly.

The second piece of information that will be important to you is the size of the particular unit. This information is located just above the symbol. In this study, we used the symbols (from smallest to largest) for Squad (• or **1**), Platoon (••• or **5**), Company (**|** or **6**), Battalion (**| |** or **7**), and Division (**XX** or **10**). In the maps scenarios that you will view, smaller units will be considered less of a threat than larger ones.

Do you have any questions about these symbols? During the study, you'll have this cue card as a reminder.

Show Map 1.
This map is characteristic of what you will see during the study. Your unit will always located in the center of the screen with other units scattered to the east and west. In addition, there will be other map elements such as cities, towns, roads, and railroads.

The third piece of information that will be important to you is the distance from the unit to your position. Of course, closer enemy units may be more of a threat than ones that are further away. I say ‘may’ here, because this threat will be influenced by the fourth piece of important information, the difficulty of terrain between units.

(*Point to different regions of contour lines*). Are you familiar with the use on contour lines on maps? (If so then, you’ll know that) A contour line connects points on the land that have the same elevation. In general, contour lines that are close together, like here, indicate a region where the terrain is steep and more difficult to traverse than a region, such as here, where the lines are further apart (and therefore relatively flat). For this study, the exact contour interval is not important, only the relative difficulty of regions on the same map.

It is important that you use both the distance information and the difficulty of terrain to determine how accessible you are for a particular unit. For instance, a smaller force that is more distant over easy (flat) terrain may be more of a threat than a larger force that is nearby over difficult terrain (*point to map*).

Do you have any questions?

The final piece of information that you will need to consider is the reliability of the information being displayed on the map. During actual combat situations, a commander may be presented with reports and information that is very unreliable versus information that is highly reliable (confirmed). For this study, the border of the symbol will note the reliability of the units. (*Show instruction image 2*) Here you can see three types of border: the solid border will denote highly reliable information (confirmed), a dashed border will denote information that is of medium reliability, and a dotted border will denote very unreliable information.

So, now you have all the required information that you will need to determine the threat of a particular unit: the type (enemy, neutral, or friendly), the size of force, the distance from your position, the difficulty of the terrain, and the reliability of the information. The overall likelihood of an attack from the east or west should be assessed based on the integrated value of all the units on each side of the map (that is, the sum threat of each unit in the east versus that of the west).

Do you have any questions so far?

I know this task seems to be quite an undertaking – you’ll be pleased to hear however that on some of the trials you will have an aide to help in your assessment. On these trials, the computer will automatically assess the battlefield and enhance only the units that are of highest threat. All other units will appear normal, as these enhanced units will pulse from low to high intensity. The purpose of this automation is to guide your attention to the most relevant units on the battlefield, perhaps saving you from having to do so yourself. We note that items that are not highlighted are not necessarily irrelevant, that is, they can pose *some* threat to you. They are simply deemed to be *less* of a threat than the highlighted units.

Do you have any questions?

(*Show Map 2*)
Here is a sample of what a battlefield may look like. This is your position in the center. As you can see, other units are distributed to the east and west of your position. You will need to assess the overall threat from each direction and then allocate 20 "units" of defensive resources to either side of your positions. For example, if you decide that the threat from the east is 50% greater than for the west, you could allocate 12 resources to the east side and 8 to the west side. There are no correct or incorrect responses here, but you should try to match the relative threat and your allocation as closely as possible. Is this clear?

A few final points that will help you as you go through the scenarios,
- You can assume that all units will approach your position on a straight (direct) path. Like so.
- All units on the map are acting independently from one another. They will not interfere with one another or impede other's progress.
- All neutral (and of course friendly) units are NO threat to yourself.

You will be viewing 56 different battlefield scenarios (including practice). Each will start with a brief instruction screen. Press any key and the trial will start. You will have up to 35 seconds to observe the map. If you are ready to respond before this time is up, press any key and you will be taken to the response screen, where you can input the number of units allocated in either box (the other will fill in automatically). After responding, you can press any key to start the next trial (whenever you are ready). On a few rare occasions we may ask you about the identity of a specific cue following the scenario.

Now I will show you some practice trials so you can get used to the task at hand.

Any questions before we begin?

Show practice block.

Any questions?

When you are ready, I will start the next segment. There are 35 trials total. The automatic aid will appear on some, but not all of the trials.
Have fun!
Show experimental blocks.
Give participant post-experimental questionnaire. Go through form with participant.

Answer questions.
Thank and remunerate participant.

Appendix J

Post-Experimental Questionnaire: Study 1

Participant_____

Please rate the following cues on their information value (i.e., how important they were) for your task:

	Not at All Informative	Slightly Informative	Moderately Informative	Very Informative	Extremely Informative
1) Size of Unit	1	2	3	4	5
2) Distance from your Unit	1	2	3	4	5
3) Difficulty of Terrain (between unit and your position)	1	2	3	4	5
4) Type of Unit (e.g., enemy, friendly)	1	2	3	4	5
5) Reliability of Information	1	2	3	4	5

6) When deciding how to allocate your defensive resources, the automation feature was:

Not at All Useful	Slightly Useful	Moderately Useful	Very Useful	Extremely Useful
1	2	3	4	5

7) Did you encounter any problems or difficulties while using the map display? If so, please describe.

8) Under what conditions would you consider using the automation feature to help guide your attention in a real battlefield situation?

9) Did the enhancements help you notice potential threats?

10) Did the enhancements interfere with your ability to allocate resources accordingly?

11) What did you like about the enhancements?

12) What did you dislike about the enhancements?

Appendix K

Experimental Scenario Blocks: Study 1

Order	1	2	3	4
Scenario # (A = automation; MP = Memory Probe; F = Failure)	34 (A)	24	34	24 (A)
	51 (A)	15	51	15 (A)
	44 (A)	11	44	11 (A)
	46 (A)	5	46	5 (A)
	4	57	4 (A)	57 (A)
	6	32 (A)	6 (A)	32
	660 (MP 1)	56 (A)	660 (MP 1, A)	56
	63	40 (A)	63 (A)	40
	42 (A)	59 (A)	42	59
	48 (A)	50 (A)	48	50
	43 (A)	14	43	14 (A)
	55 (A)	35	55	35 (A)
	13	23	13 (A)	23 (A)
	21	17	21 (A)	17 (A)
	19	770 (MP 2, A)	19 (A)	770 (MP 2)
	26	36 (A)	26 (A)	36
	49 (A)	30 (A)	49	30
	39 (A)	60 (A)	39	60
	33 (A)	1	33	1 (A)
	47 (A)	16	47	16 (A)
	18	61	18 (A)	61 (A)
	12	28	12 (A)	28 (A)
	3	38 (A)	3 (A)	38
	22	64 (A)	22 (A)	64
	7 (A)	53 (A)	7	53
	53 (A)	7 (A)	53	7
	64 (A)	22	64	22 (A)
	37 (A)	2	37	2 (A)
	25	12	25 (A)	12 (A)
	62	18	62 (A)	18 (A)
	8	45 (A)	8 (A)	45
	1	29 (A)	1 (A)	29
	60 (A)	39 (A)	60	39
	31 (A)	49 (A)	31	49
	41 (A)	26	41	26 (A)
	770 (MP 2, A)	20	770 (MP 2)	20 (A)
	17	21	17 (A)	21 (A)
	23	10	23 (A)	10 (A)
	27	54 (A)	27 (A)	54
	9	52 (A)	9 (A)	52
	50 (A)	48 (A)	50	48
	58 (A)	42 (A)	58	42
	40 (A)	63	40	63 (A)
	56 (A)	660 (MP 1)	56	660 (MP 1, A)
	32 (A)	6	32	6 (A)
	57	4	57 (A)	4 (A)
	5	46 (A)	5 (A)	46
	11	44 (A)	11 (A)	44
	15	51 (A)	15 (A)	51
	24	34 (A)	24 (A)	34
	990 (F, A)	990 (F, A)	990 (F, A)	990 (F, A)

Appendix L

Participant Responses: Study 1

Under what conditions would you consider using the automation feature to help guide your attention?
- under time pressure (7)
- at a very high level of command (Brigade or higher)
- night and low visibility or dense vegetation
- if it was very reliable information (i.e., truly told me where the strong units were located)
- best used when moderate amounts of time is available so that the most likely area of first contact would be covered. It would be less effective when there is little time to plan.
- when there are many different enemy units in various locations, clutter (3)
- in situations where terrain or enemy locations is unclear
- when reliability of all information on screen is moderate to low and distance and terrain are similar for all enemy units
- maybe to consult with once I made a decision (to see what the automation would have suggested)
- when trying to quickly decided which area had either a larger force or relatively easy terrain to cross to reach my position

Did the enhancements help you notice potential threats?
- in some cases (3)
- yes (15)
- not really. The enemy units (diamond shape) were sufficiently distinct. (2)

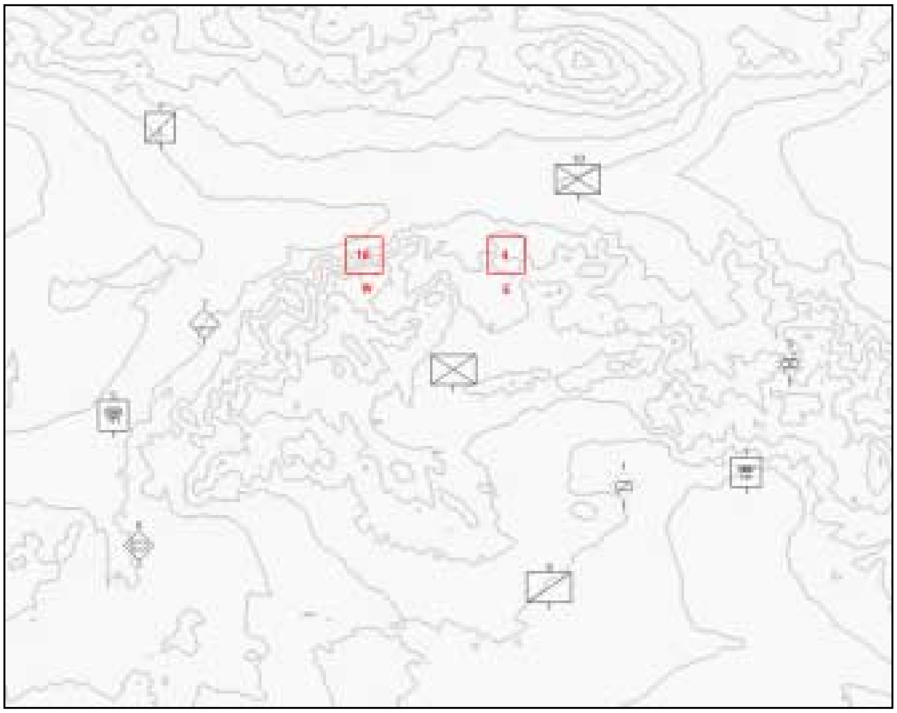
Did the enhancements interfere with your ability to allocate resources accordingly?
- no (13)
- I tried not to let that happen. There were a few times when the enhancement did not highlight a large enemy unit so I allocated ‘against’ the enhancements.
- yes, sometimes a squad level enemy was flashing, drawing my attention when I should have been paying more attention to larger units further away.
- somewhat. I found it hard to focus on other enemy elements. (2)
- I sometimes ignored it

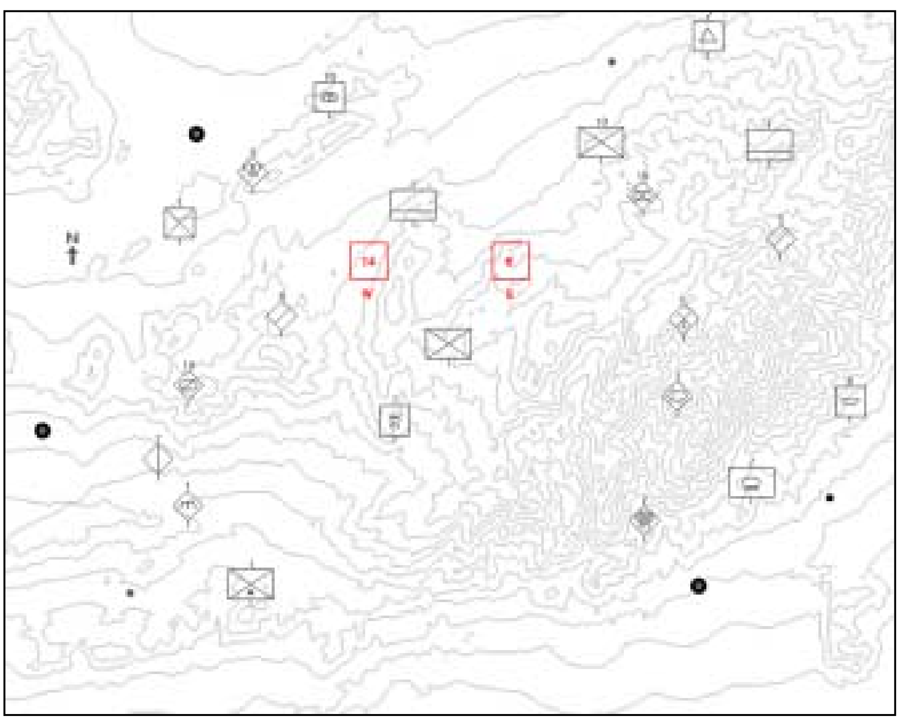
What did you like about the enhancements?
- it's a quick way to determine potential threats (2)
- it did help discriminate enemy forces from the extra info (friendly / neutral) (3)
- drew my attention to potential threats immediately (2)
- made me more alert, provided a more interactive situation
- allowed me to focus on what were the most important areas to allocate
- highlighted larger forces that were further away that might otherwise have been ignored
- it gave you info about the importance (threat) of a particular unit
- it appears to be an easy way to organize a great deal of information
- it identified possible enemy threats that were more dangerous
- most times, it pointed out units which had easier terrain
- when I had little time to make a decision, it helped me focus in on something
- helped allocate attention where needed in a cluttered display
- it reduced my scan time
- provided a starting point for situation assessment
- helped me recall where I had seen enemy units

What did you dislike about the enhancements?
- they could have a tendency to distract your attention from a potential hazard
- it made it more difficult as a person to attempt to consider the enemy forces that were not highlighted
- sometimes they were overwhelming when everything on the screen was enhanced
- I sometimes ignored items not enhanced (2)
- sometimes it assess threats differently than I would have (3)
- sometimes there were too many. Also, it was difficult to decide how low reliability and flashing interacted.
- I would have liked a color display (i.e., red indicating greater risk)
- not reliable enough (3)
- size of enemy unit sometimes did not seem to convey the amount of threat that was present with other flashing units
- distracting (3)
- I don't know if I would always trust a computer
- did not understand how computer decided on what was the biggest threats (2)
- imposed a secondary task (i.e., assessing computer's judgment in addition to my own)
- I may have relied on it too much when making my decision
- I spent time trying to figure out why some enemies were enhanced

Appendix M

Sample battlefield scenarios with automated aid: Study 2





Appendix N

Verbal Protocol: Study 2

General Instructions

Provide participant with informed consent.

Thank you very much for participating in this study. It should take approximately 45 minutes to complete. I would like to remind you that you are free to withdraw from this study at any time. Please look through the informed consent.

Participant reads / signs forms.

Do you have any questions?

To begin with, I will ask you to fill out this brief questionnaire. It will ask you a few background questions.

Participant fills out questionnaire.

Today I will show you some electronic maps of battlefields. Your unit is positioned in the center of the map. Your task will be to observe the other units in the area and decide from which direction an enemy attack is more likely. Based on this decision, you will allocate your defensive resources accordingly.

I'd like to over some of the things you'll need to pay attention to when making your assessment. We have attempted to match the symbols to the standard military ones you may be familiar with. Here is a small sample of symbols used here (*Show instruction image 1*).

First of all, note the color and shape of the symbols. Enemy units are marked by DIAMONDS, neutral units are marked by SQUARES, and friendly units are marked by RECTANGLES. You'll note that inside of each shape is a unit type (e.g., light infantry or engineers – combat dismounted). For the purposes of this study, this unit type will not be important and can be ignored, only whether or not the unit is an enemy, neutral, or friendly.

The second piece of information that will be important to you is the size of the particular unit. This information is located just above the symbol. In this study, we used the symbols (from smallest to largest) for Squad (**1**), Platoon (**5**), Company (**6**), Battalion (**7**), and Division (**10**). In the maps scenarios that you will view, smaller units will be considered less of a threat than larger ones.

Do you have any questions about these symbols? During the study, you'll have this cue card as a reminder.

Show Map 1.

This map is characteristic of what you will see during the study. Your unit will always located in the center of the screen with other units scattered to the east and west. In addition, there will be other map elements such as cities, towns, roads, and railroads.

The third piece of information that will be important to you is the distance from the unit to your position. Of course, closer enemy units may be more of a threat than ones that are further away. I say 'may' here, because this threat will be influenced by the fourth piece of important information, the difficulty of terrain between units.

(*Point to different regions of contour lines*). Are you familiar with the use on contour lines on maps? (If so then, you'll know that) A contour line connects points on the land that have the same elevation. In general, contour lines that are close together, like here, indicate a region where the terrain is steep and more difficult to traverse than a region, such as here, where the lines are further apart (and therefore relatively flat). For this study, the exact contour interval is not important, only the relative difficulty of regions on the same map.

It is important that you use both the distance information and the difficulty of terrain to determine how accessible you are for a particular unit. For instance, a smaller force that is more distant over easy (flat) terrain may be more of a threat than a larger force that is nearby over difficult terrain (*point to map*).

Do you have any questions?

The final piece of information that you will need to consider is the reliability of the information being displayed on the map. During actual combat situations, a commander may be presented with reports and information that is very unreliable versus information that is highly reliable (confirmed). For this study, the border of the symbol will note the reliability of the units. (*Show instruction image 2*) Here you can see three types of border: the solid border will denote highly reliable information (confirmed), a dashed border will denote information that is of medium reliability, and a dotted border will denote very unreliable information.

(*Show instruction image 3*)
So, now you have all the required information that you will need to determine the threat of a particular unit: the type (enemy, neutral, or friendly), the size of force, the distance from your position, the difficulty of the terrain, and the reliability of the information. The overall likelihood of an attack from the east or west should be assessed based on the integrated value of all the units on each side of the map (that is, the sum threat of each unit in the east versus that of the west).

Do you have any questions so far?

(*Show Map 2*)

Here is a sample of what a battlefield may look like. This is your position in the center. As you can see, other units are distributed to the east and west of your position. You will need to assess the overall threat from each direction and then allocate 20 "units" of defensive resources to either side of your positions. For example, if you decide that the threat from the east is 50% greater than for the west, you could allocate 12 resources to the east side and 8 to the west side. There are no correct or incorrect responses here, but you should try to match the relative threat and your allocation as closely as possible. Is this clear?

I know this task seems to be quite an undertaking – you'll be pleased to hear however that on some of the trials you will have an aide to help in your assessment. On these trials, the computer will automatically assess the battlefield and suggest an appropriate allocation of defensive resources. This is only a suggestion, you are free to allocate your defenses however YOU deem appropriate. This automation is, in general, highly reliable however is not perfect.

Do you have any questions?

A few final points that will help you as you go through the scenarios,
- You can assume that all units will approach your position on a straight (direct) path. Like so.
- All units on the map are acting independently from one another. They will not interfere with one another or impede other's progress.
- All neutral (and of course friendly) units are NO threat to yourself.

You will be viewing 48 different battlefield scenarios (including practice). Each will start with a brief instruction screen. Press any key and the trial will start. You will have up to 25 seconds to observe the map. If you are ready to respond before this time is up, press any key and you will be taken to the response screen, where you can input the number of units allocated in either box (the other will fill in automatically). After responding, you can press any key to start the next trial (whenever you are ready). On a few rare occasions we may ask you about the identity of a specific cue following the scenario.

Now I will show you some practice trials so you can get used to the task at hand.

Any questions before we begin?

Show practice block.

Any questions?

When you are ready, I will start the next segment. There are 43 trials total. The automatic aide will appear on some, but not all of the trials.

Have fun!
Show experimental blocks.

Give participant post-experimental questionnaire. Go through form with participant.
Answer questions.
Thank and remunerate participant.

Appendix O

Pre-Experimental Questionnaire: Study 2

Participant _____

1. Age _________
2. Gender _______
3. Do you have normal or corrected-to-normal vision? Yes No

Please rate your experience with (or understanding of) the following:

	None		Moderate		Very High
4. Contour maps	1	2	3	4	5
5. Military symbology (e.g., unit size, type)	1	2	3	4	5

Appendix P

Post-Experimental Questionnaire: Study 2

Participant_____

Please rate the following cues on their information value (i.e., how important they were) for your task:

	Not at All Informative	Slightly Informative	Moderately Informative	Very Informative	Extremely Informative
1) Size of Unit	1	2	3	4	5
2) Distance from your Unit	1	2	3	4	5
3) Difficulty of Terrain (between unit and your position)	1	2	3	4	5
4) Type of Unit (e.g., enemy, friendly)	1	2	3	4	5
5) Reliability of Information	1	2	3	4	5

6) When deciding how to allocate your defensive resources, the automation feature was:

Not at All Useful	Slightly Useful	Moderately Useful	Very Useful	Extremely Useful
1	2	3	4	5

7) Did you encounter any problems or difficulties while using the map display? If so, please describe.

8) Under what conditions would you consider using the automation feature to help in battlefield decision making?

9) What did you like about the automation?

10) What did you dislike about the automation?

Appendix Q

Participant Responses: Study 2

Under what conditions would you consider using the automation feature to help in battlefield decision-making?
- when knowing its reliability and how it makes its decisions
- under time pressure (4)
- low threat conditions
- with supervision / review of experienced operator
- if shown to be very reliable in generating appropriate decisions
- if I was actually in battle, I'd almost have to trust the computer's risk assessment. For this particular task, I relied more on the computer when the cues were in conflict (i.e., bigger but more distant or closer but less reliable)
- if I'm unsure and want a second opinion (to validate my impression of the situation)
- conditions of uncertainty (3)
- in complex situations

What did you like about the automation?
- helped focus attention to the side with more enemies (2)
- made it easier to go from those numbers to double-check using the map (2)
- fairly accurate
- usually appropriate decision suggested – used as a "ballpark" figure to assess threat (2)
- it gave me a baseline for allocating defense units
- offered a second opinion for difficult decisions (2)
- helped shape my assessment of the situation
- could use it to figure out how the computer weighed the various cues (2)

What did you dislike about the automation?
- didn't know what the decisions were based on (2)
- sometimes, I didn't agree with the numbers the computer generated (but it's hard to contradict a computer)
- it was sometimes more conservative and sometimes more extreme (than I)
- did not seem to take enemy distance into account
- seemed inaccurate in some instances (4)
- potential to bias operator's assessment
- it lowered my confidence and influenced my choices more than I would have liked
- another distraction (though more useful)
- uncertain of its reliability (after it differed from my own decision)
- did not provide any reasoning behind the suggestion
- often disagreed with it and second guessed myself as a result (2)